Aus dem Bereich Klinische Bioinformatik
Klinische Medizin
der Medizinischen Fakultät
der Universität des Saarlandes, Homburg/Saar

Resources for the analysis of bacterial and microbial genomic data with a focus on antibiotic resistance

Dissertation zur Erlangung des Grades eines Doktors
der Naturwissenschaften der Medizinischen Fakultät

der UNIVERSITÄT DES SAARLANDES
2019

vorgelegt von Valentina Galata
geb. am 31.05.1988 in Rostow am Don, Russland

Tag der Promotion:	18. Juni 2020
Dekan:	Univ.-Prof. Dr. med. Michael D. Menger
Berichterstatter:	Prof. Dr. Andreas Keller
	Prof. Dr. Eckart Meese
	Prof. Dr. Karsten Becker

"All models are wrong, but some are useful."

— George E. P. Box

Abstract

Antibiotics are drugs which inhibit the growth of bacterial cells. Their discovery was one of the most significant achievements in medicine: it allowed the development of successful treatment options for severe bacterial infections, which has helped to significantly increase our life expectancy. However, bacteria have the ability to adapt to changing environmental conditions through genetic modifications, and can, therefore, become resistant to an antibiotic. Extensive use of antibiotics promotes the development of antibiotic resistance and, since some genetic factors can be exchanged between the cells, emergence of new resistance mechanisms and their spread have become a serious global problem.

Counteractive measures have been initiated, focusing on the different factors contributing to the antibiotic resistance crisis. These include the study of bacterial isolates and complete microbial communities using whole-genome sequencing (WGS) data. In both cases, there are specific challenges and requirements for different analytical approaches. The goal of the present thesis was the implementation of multiple resources which should facilitate further microbiological studies, with a focus on bacteria and antibiotic resistance. The main project, GEAR-base, included an analysis of WGS and resistance data of around eleven thousand bacterial clinical isolates covering the main human pathogens and antibiotics from different drug classes. The dataset consisted of WGS data, antibiotic susceptibility profiles and meta-information, along with additional taxonomic characterization of a sample subset. The analysis of this isolate collection allowed for the identification of bacterial species demonstrating increasing resistance rates, to construct species pan-genomes from the *de novo* assembled genomes, and to link gene presence or absence to the available antibiotic resistance profiles. The generated data and results were made available through the online resource GEAR-base. This resource provides access to the resistance information and genomic data, and implements functionality to compare submitted genes or genomes to the data included in the resource.

In microbial community studies, the metagenome obtained through WGS is analyzed to determine its taxonomic composition. For this task, genomic sequences are clustered, or binned, to represent sequences belonging to specific organisms or closely-related organism groups. BusyBee Web was developed to provide an automatic binning pipeline using frequencies of *k*-mers (subsequences of length *k*)

and bootstrapped supervised clustering. It also includes further data annotation, such as taxonomic classification of the input sequences, presence of know resistance factors, and bin quality.

Plasmids, extra-chromosomal DNA molecules found in some bacteria, play an important role in antibiotic resistance spread. As the classification of sequences from WGS data as chromosomal or plasmid-derived is challenging, demonstrated by evaluating four tools implementing three different approaches, having a reference dataset to detect the plasmids which are already known is therefore desirable. To this end, an online resource for complete bacterial plasmids (PLSDB) was implemented.

In summary, the herein described online resources represent valuable datasets and/or tools for the analysis of microbial genomic data and, especially, bacterial pathogens and antibiotic resistance.

Zusammenfassung

Antibiotika sind Medikamente, die das Wachstum von Bakterienzellen hemmen. Ihre Entdeckung war eine der bedeutendsten Leistungen der Medizin: Es erlaubte die Entwicklung von erfolgreichen Behandlungsmöglichkeiten von schwerwiegenden bakteriellen Infektionen, was geholfen hat, unsere Lebenserwartung zu erhöhen. Allerdings sind Bakterien in der Lage sich den wechselnden Umweltbedingungen anzupassen und können dadurch resistent gegen ein Antibiotikum werden. Der extensive Gebrauch von Antibiotika fördert die Entwicklung von Antibiotikaresistenzen und, da einige genetische Faktoren zwischen den Zellen ausgetauscht werden können, sind das Auftauchen von neuen Resistenzmechanismen und deren Verbreitung zu einem seriösen globalen Problem geworden.

Gegenmaßnahmen wurden ergriffen, die sich auf die verschiedenen Faktoren fokussieren, die zur Antibiotikaresistenzkrise beitragen. Diese umfassen Studien von bakteriellen Isolaten und ganzen Mikrobengemeinschaften mithilfe von Gesamt-Genom-Sequenzierung (GGS). In beiden Fällen gibt es spezifische Herausforderungen und Bedürfnisse für verschiedene analytische Methoden. Das Ziel dieser Dissertation war die Implementierung von mehreren Ressourcen, die weitere mikrobielle Studien erleichtern sollen und einen Fokus auf Bakterien und Antibiotikaresistenz haben. Das Hauptprojekt, GEAR-base, beinhaltete eine Analyse von GGS- und Resistenzdaten von ungefähr elftausend klinischen Bakterienisolaten und umfasste die wichtigen menschlichen Pathogene und Antibiotika aus verschiedenen Medikamentenklassen. Neben den GGS-Daten, Empfindlichkeitsprofilen für die Antibiotika und Metainformation, beinhaltete der Datensatz zusätzliche taxonomische Charakterisierung von einer Teilmenge der Proben. Die Analyse dieser Sammlung an Isolaten erlaubte die Identifizierung von Spezies mit ansteigenden Resistenzraten, die Konstruktion von den Spezies-Pan-Genomen aus den *de novo* assemblierten Genomen und die Verknüpfung vom Vorhandensein oder Fehlen von Genen mit den Antibiotikaresistenzprofilen. Die generierten Daten und Ergebnisse wurden durch die Online-Ressource GEAR-base bereitgestellt. Diese Ressource bietet Zugang zur Resistenzinformation und den gesammelten genomischen Daten und implementiert Funktionen zum Vergleich von hochgeladenen Genen oder Genomen zu den Daten, die in der Ressource enthalten sind.

In den Studien von Mikrobengemeinschaften wird das durch GGS erhaltene Metagenom analysiert, um seine taxonomische Zusam-

mensetzung zu bestimmen. Dafür werden die genomischen Sequenzen in sogenannte Bins gruppiert (Binning), die die Zugehörigkeit von den Sequenzen zu bestimmten Organismen oder zu Gruppen von nah verwandten Organismen repräsentieren. BusyBee Web wurde entwickelt, um eine automatische Binning-Pipeline anzubieten, die die Häufigkeitsprofile von *k*-meren (Teilsequenzen der Länge *k*) und eine auf dem Bootstrap-Verfahren basierte Methode für die Gruppierung der Sequenzen nutzt. Zusätzlich wird eine Annotation der Daten durchgeführt, wie die taxonomische Klassifizierung der hochgeladenen Sequenzen, das Vorhandensein von bekannten Resistenzfaktoren und die Qualität der Bins.

Plasmide, DNA-Moleküle, die zusätzlich zum Chromosom in einigen Bakterien vorhanden sind, spielen eine wichtige Rolle in der Verbreitung von Antibiotikaresistenzen. Die Klassifizierung von Sequenzen aus der GGS als von einem Chromosom oder einem Plasmid stammend ist herausfordernd, wie es in einer Evaluation von vier Tools, die drei verschiedene Ansätze implementieren, demonstriert wurde. Deshalb ist das Vorhandensein von einem Referenzdatensatz, um schon bekannte Plasmide zu detektieren, sehr wünschenswert. Zu diesem Zweck wurde eine Online-Ressource von vollständigen bakteriellen Plasmiden implementiert (PLSDB).

Die hier beschriebenen Online-Ressourcen stellen nützliche Datensätze und/oder Werkzeuge dar, die für die Analyse von mikrobiellen genomischen Daten, insbesondere von bakteriellen Pathogenen und Antibiotikaresistenzen, eingesetzt werden können.

Scientific papers

This is a cumulative thesis based on the following published papers. The publications included herein are identical to the published versions.

- **V. Galata**, C. Backes, C. C. Laczny, G. Hemmrich-Stanisak, H. Li, L. Smoot, A. E. Posch, S. Schmolke, M. Bischoff, L. von Muller, A. Plum, A. Franke, and A. Keller. Comparing genome versus proteome-based identification of clinical bacterial isolates. *Brief. Bioinformatics*, 19(3):495–505, May 2018. [1]

- **V. Galata**, C. C. Laczny, C. Backes, G. Hemmrich-Stanisak, S. Schmolke, A. Franke, E. Meese, M. Herrmann, L. von Muller, A. Plum, R. Muller, C. Stahler, A. E. Posch, and A. Keller. Integrating Culture-based Antibiotic Resistance Profiles with Whole-genome Sequencing Data for 11,087 Clinical Isolates. *Genomics Proteomics Bioinformatics*, 17(2):169–182, Apr 2019. [2]

- C. C. Laczny, C. Kiefer, **V. Galata**, T. Fehlmann, C. Backes, and A. Keller. BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res.* , 45(W1):W171–W179, Jul 2017. [3]

- C. C. Laczny, **V. Galata**, A. Plum, A. E. Posch, and A. Keller. Assessing the heterogeneity of in silico plasmid predictions based on whole-genome-sequenced clinical isolates. *Brief. Bioinformatics*, 20(3):857–865, May 2019. [4]

- **V. Galata**, T. Fehlmann, C. Backes, and A. Keller. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.* , 47 (D1):D195–D202, Jan 2019. [5]

Contents

List of Figures

List of Tables

Abbreviations

AST antibiotic susceptibility testing 42

BLP Braun's (or murein) lipoprotein 26

CARD Comprehensive Antibiotic Resistance Database 109

CDC Centers for Disease Control and Prevention 36, 41

DHFR dihydrofolate reductase 29, 31

DHPS dihydropteroat synthase 29

ECDC European Centre for Disease Prevention and Control 41

GI genomic island 35

GyrA subunit A of DNA gyrase 27, 28, 30

HGT horizontal gene transfer 33, 34, *Glossary:* HGT

IS insertion sequence 34, 35

LPS lipopolysaccharide 26

MDR multi-drug resistant 29, *Glossary:* MDR

MIC minimal inhibitory concentration 29, 42, 107, *Glossary:* MIC

MLST multi-locus sequence typing 42

MS mass spectrometry 40, 47, 107

NCBI National Center for Biotechnology Information 32, 42

NGS next-generation sequencing 44

OTU operational taxonomic unit 44

ParC subunit A of topoisomerase IV 27, 28, 30

PBP penicillin-binding protein 26, 31

PDR pan-drug resistant 29, *Glossary:* PDR

REP repetitive extragenic palindromic sequence 34

VFDB Virulence Factor Database 109

WGS whole-genome sequencing 40–42, 44, 45, 47, 107, 109

WHO World Health Organization 36, 41

XDR extensively resistant 29, *Glossary:* XDR

Glossary

GC content Proportion of guanine/cytosine bases in a genome 32, 35

HGT Transfer of genetic material between unrelated genomes 33

MDR Multi-drug resistant – resistant to several antimicrobials or drug classes 29

metagenome Collection of genes and genomes of all the organisms of a microbiota 21

MIC Minimal concentration of antibiotic that inhibits visible growth 29

microbiome The metagenome of a microbiota and the colonized environment 21, 44

microbiota Community of microbes found in an environment 21, 38, 42, 44, 45

pan-genome Collection of all the genes found in a group of organisms 32

PDR Pan-drug resistant – resistant to all antibiotics 29

XDR Extensively resistant – a more extreme case of multi-drug resistance 29

1
The threat of antibiotic resistance

1.1 Bacteria

All living organisms on Earth are classified into two major groups — prokaryotes and eukaryotes [6]. The latter are organisms with cells containing organelles enclosed by a membran such as nucleous or mitochondria [6]. These organisms include fungi, protozoa, eukaryotic algae, plants, and animals [6]. Organisms lacking internal membranes and thus not possessing such organelles are termed prokaryotes [6]. Prokaryotes are further divided into archaea and bacteria, forming together with the Eucarya (eukaryotes) the "three domains of life" (Figure 1.1) — a concept proposed by Woese *et al.* in 1990 [7]. In general, organisms which cannot be seen by the naked human eye are also grouped together under the term microbes (or microorganisms) which include, besides viruses, fungi and protozoa, all archaea and bacteria [8].

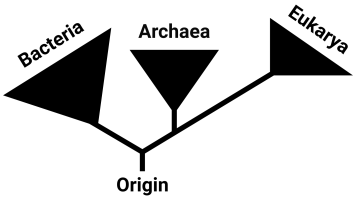


Figure 1.1: Three domains of life, adapted from [7].

Occurring almost everywhere, with an estimated total number of 9.2×10^{29} to 31.7×10^{29} of prokaryotic cells on Earth [9], microbes are involved in many fundamental processes in ecological systems. Bacteria are present in virtually all ecological environments, including such extreme habitats as sea ice, hot springs, and hypersaline and alkaline lakes [10–12]. Many bacterial species interact with eukaryotic cells establishing commensal, symbiotic, mutualistic, parasitic or pathogenic relationships with their host [13]. Some bacteria are known to reside within the host cell as endosymbionts [13] while others are found outside of the host cells, e.g. on outer or inner organ surfaces such as skin [14] and colon [15]. The symbiont-host relationship can be harmless (commensalism) and, additionally, beneficial for one (symbiosis) or both (mutualism) organisms. In contrast, a parasitic or pathogenic relationship is characterized by detrimental consequences for the host [16].

An environment is generally not colonized by a single bacterial organism but by bacterial communities (microbiota) [17; 18]. The collection of genes and genomes of all the associated microorganisms is referred to as the metagenome [17]. Together, the metagenome of a microbiota and the colonized environment constitute the microbiome [17]. In humans, bacterial communities can be found on different parts of the body, including our skin, oral cavity, airways and gastrointestinal tract [19]. According to a recent estimation, the ratio of bacterial and human cells in a "reference man" is $1.26\bar{6} : 1$ with approximately 3.8×10^{13} bacterial and 3.0×10^{13} human cells [20]. We maintain a mutualistic, symbiotic relationship with our microbes [21]. While we provide them with space and nutrients, the microbes play an important role in digestion, immune system development, and protect us from pathogens [22]. Therefore, the integrity of the microbial community within us has a great impact upon our health.

1.1.1 *Discovery and classification of bacteria*

In 1676, Antony Leeuwenhoek reported his discovery of small living organisms in a water sample using a microscope of his own design [23]. However, the theory that these microorganisms are responsible for many diseases (the "germ theory of disease") was only developed in the nineteenth century [24]. The "golden age of microbiology" started after the discovery of *Bacillus anthracis*, a pathogen causing anthrax in livestock and humans, by Robert Koch in 1876 [24]. Over the next 30 years, many other bacterial pathogens were identified, including those responsible for tuberculosis, cholera, and plague [24]. Subsequent discoveries in molecular biology and the development of new technologies deepened our knowledge and understanding of bacterial infections and bacteria in general. Advances in genome sequencing had a profound impact on the study of bacterial organisms. Today, the sequencing of specific genomic regions or complete genomes of single organisms or communities is a fundamental tool for the identification, characterization, and analysis of microbes [25].

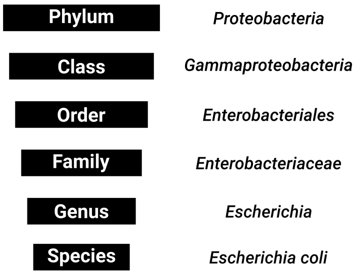


Figure 1.2: Main taxonomic ranks and their hierarchy (left) with an example for the *Escherichia coli* species (right). There are also other ranks above (kingdom and domain), below (e.g. sub-species), and between (e.g. sub-order) those shown herein.

In order to identify and characterize bacteria, a taxonomy system is required. Taxonomies include guidelines for nomenclature, classification criteria, and a hierarchical representation of organism groups using multiple taxonomic ranks (Figure 1.2). The first bacterial taxonomy system, produced by Ferdinand Cohn in 1872, was based on morphological characteristics of the cells (e.g. cell size and shape) [26]. A different system, considering the physiological properties of bacteria, was proposed by Orla-Jensen in 1909 [26]. Various classification systems were then developed using modified versions of Cohn's morphology-based approach [26]. The discovery of DNA and the development of sequencing technologies brought substantial changes: towards the end of the 20th century, DNA-DNA hybridization and analysis of ribosomal RNA (rRNA) gene sequences (in particular the 16S rRNA gene) have been applied to delineate bacterial species [27–29] (see also 1.4.5). The current taxonomic classification relies on a polyphasic approach taking into account different traits including phenotypic and genomic characteristics [30].

It is important to note that taxonomic descriptions are not fixed, but constantly revised [31]. This results in multiple synonymous names, the merging and splitting of existing taxa, and changes in the placement of the bacterial organisms within the taxonomic hierarchy. The taxonomic classification approaches currently in use are not optimal [32] and, as the number of completely sequenced bacterial genomes increases, new classification procedures and criteria have been proposed. These include the estimation of genomic relatedness using average nucleotide identity (ANI) [33] and creating new groupings using ubiquitous single-copy proteins found in all bacterial genomes [34].

1.1.2 Bacterial infections

Besides commensal microbes colonizing our bodies, there are also pathogenic bacteria responsible for many, often deadly, diseases which pose a major risk to public health. As mentioned above, the discovery of the causative organisms of bacterial infections began towards the end of the nineteenth century (Table 1.1). Some infectious diseases, such as cholera and plague, had a devastating impact on human populations through epidemics and pandemics [35, pp. 969–979]. For example, *Yersinia pestis* caused several plague pandemics, with the most recent one resulting in over 10 million deaths [36].

Disease	**Pathogen**
Anthrax	*Bacillus anthracis*
Suppuration	*Staphylococcus*
Gonorrhea	*Neisseria gonorrhoeae*
Typhoid fever	*Salmonella typhi*
Suppuration	*Streptococcus*
Tuberculosis	*Mycobacterium tuberculosis*
Cholera	*Vibrio cholerae*
Diphtheria	*Corynebacterium diptheriae*
Tetanus	*Clostridium tetani*
Diarrhea	*Escherichia coli*
Pneumonia	*Streptococcus pneumoniae*
Meningitis	*Neisseria meningtidis*
Food poisoning	*Salmonella enteritidis*
Gas gangrene	*Clostridium perfringens*
Plague	*Yersinia pestis*
Botulism	*Clostridium botulinum*
Dysentery	*Shigella dysenteriae*
Paratyphoid	*Salmonella paratyphi*
Syphilis	*Treponema pallidum*
Whooping cough	*Bordtella pertussis*

Table 1.1: Main bacterial pathogens discovered during the "golden age of microbiology", adapted from [37].

Today, socomial (community-acquired) and nosocomial (hospital-acquired) bacterial infections still have a large impact on morbidity (presence of a medical condition) and mortality (death) rates [38; 39]. Among the most frequent infections are pneumonia [40], tuberculosis [41], urinary tract infections [42], food-borne infections [41], and infections associated with medical interventions [43].

For some bacterial diseases, preventive measures, such as improved sanitation and hygiene, and the development of vaccines have helped to decrease the number of infections. Today, vaccination against diphtheria, tetanus, and other preventable diseases are included into the vaccination recommendations for the EU [44] and the US [45]. Nevertheless, if an infection occurs and the host's immune system is unable to defeat it, effective treatment is required to eliminate the pathogens.

1.2 *Antibiotics*

Antibiotics are drugs which inhibit the growth of bacterial cells, and occur naturally produced by some fungi and bacteria [46]. The first antibiotic, penicillin, was discovered by Alexander Fleming in 1928 [47]. Since then, a broad range of antibiotic classes have ben discovered and introduced (Figure 1.3), particularly between 1940 and 1960, a period referred to as the "golden age of antibiotics research" [48]. However, the number of new drug classes reported since 1990 is significantly smaller than in previous years.

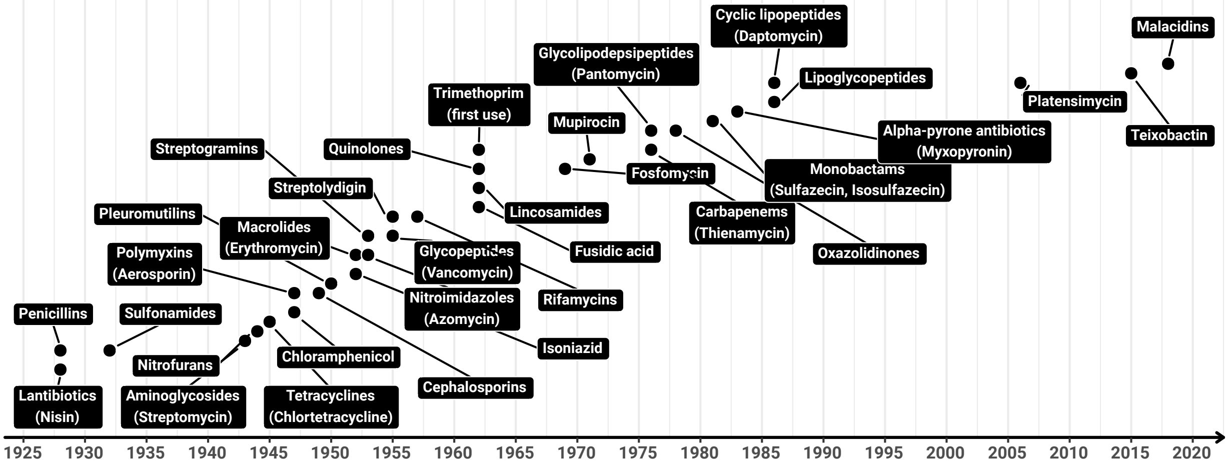


Figure 1.3: Antibiotics timeline: the (approximate) year of discovery, reporting or introduction of the major drug classes or its representatives, and antibiotics which do not belong to a specific class. The placement of the points and labels along the vertical axis is for readability purpose only. The information depicted was compiled from [49–72].

Antibiotics can be divided into categories based on their modes of action. The major drug classes include inhibitors of cell wall synthesis, compounds interacting with the outer and/or cytoplasmic membrane, inhibitors of DNA and RNA synthesis, inhibitors of protein synthesis, and inhibitors of the folic acid metabolism (Figure 1.7). Antibiotics can further be differentiated based on their chemical structure (e.g. the β-lactams which all have a β-lactam ring system [73]) and application spectrum (i.e. specific bacterial groups or even single species). In the following, an overview of antibiotic targets is given by describing the relevant cell components and processes, and how these are affected by antibiotics.

1.2.1 *Bacterial cell structure and antibiotic targets*

Bacterial cell size ranges from 0.3 µm to 750 µm [74], they demonstrate a variety of shapes including spheres, rods and spirals (Figure 1.4) [75], and form groups creating aggregates of different structures (e.g. pairs, chains and clusters, Figure 1.5) [76, ch. 2]. In nature, bacteria often form biofilms, which is a consortium of bacteria in a multi-layered structure attached to a surface and enclosed by a matrix [77].

Bacteria have a relatively simple cell composition when compared to Eukarya (Figure 1.7). On the surface, some bacterial cells possess flagella — hair-like structures used for motility with a species-characteristic distribution which can serve for organism identification [76, ch. 2]. There are also other surface filaments called pili or fimbriae which are involved in adhesion, co-aggregation, and cell-to-cell contact (sex pili) [78]. A bacterial cell is enclosed by multiple layers including, in some cases, a capsule and/or a gel layer, a cell wall, and a cytoplasmic membrane [76, ch. 2].

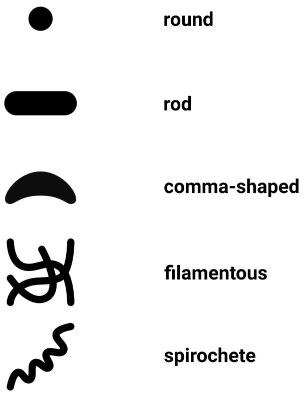


Figure 1.4: Schematic illustration of some bacterial cell shapes, adapted from [35, p. 60] and [76, ch. 2].

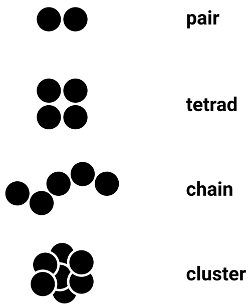


Figure 1.5: Schematic illustration of some aggregates, adapted from [76, ch. 2].

The interior of the cell — the cytoplasm or cytosol — contains a high number of ribosomes [76, ch. 2]. Often, there are also inclusions used as storage for specific compounds (e.g. glycogen) [35, p. 92]. Some bacteria which live in seas and lakes have gas vesicles, which allow them to move up and down in the water [35, p. 93]. Though bacteria do not have the organelles found in eukaryotic cells, some bacterial species possess microcompartments which could be seen as their functional analogues [79]. These cellular structures, enclosed by a protein-membrane separating them from the cytosol, have diverse functions and are linked, in some pathogenic bacteria, to virulence which provides an advantage in host colonization [79].

Bacteria generally have one circular chromosome, but some bacterial species have multiple [80] and/or linear chromosomes [81]. The chromosome is localized in a pseudo-compartment in the cytoplasm, the nucleoid [82], and is densely compacted demonstrating multiple levels of organization: the DNA-binding nucleoid-associated proteins affect the folding of the DNA and also gene expression [83]; there are multiple domains of negatively supercoiled loops of 20 kbp to 100 kbp separated from the rest of the DNA molecule by domain barriers and maintained by topoisomerases [84; 85]; DNA interactions are preferably localized within macrodomains which encompass large DNA regions (in the order of megabases) [86]. The organization of the chromosome is dynamic and changes with respect to the current state of the bacterial cell [82].

Cell wall and cytoplasmic membrane Bacterial cells are enclosed by a cell wall whose primary function is protection against osmotic lysis and maintenance of the cell shape [87]. Moreover, nearly all bacteria can be divided into two categories, Gram-positive and Gram-negative, based on the staining procedure developed by Hans Christian Gram [76, ch. 2].

The cell walls of Gram-positive bacteria have a thick multilayer shell (20 nm to 80 nm) composed of peptidoglycan (Figure 1.6) [87]. Peptidoglycan, or murein, consists of two linked sugar derivatives (N-acetylglycosamine and N-acetylmuramic acid), and amino acids attached to the second sugar [35, pp. 70–71]. A peptidoglycan layer has a repeating structure formed by the linked sugars [35, pp. 70–71]. The shell is formed of multiple peptidoglycan layers which are cross-linked through the peptide chains [35, pp. 70–71], and there are various polymers associated with this structure, including (lipo)teichoic acids [76, ch. 2]. The inner (cytoplasmic) membrane is a phospholipid bilayer which also contains some membrane-bound molecules;

it plays a significant role in active transport and biosynthesis in both Gram-positive and Gram-negative bacteria [76, ch. 2].

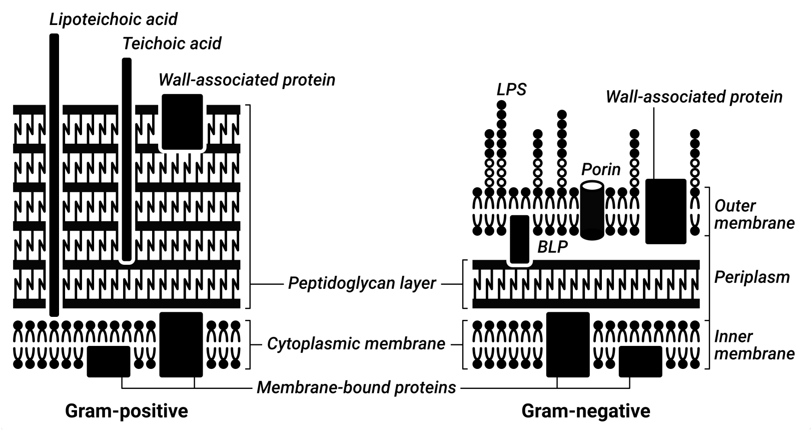


Figure 1.6: Cell wall structure of Gram-positive and Gram-negative bacteria, adapted from [35, pp. 74, 77].

Gram-negative bacteria have a thinner peptidoglycan layer (1 nm to 7 nm) located between an outer membrane and an inner membrane (Figure 1.6) [87]. Braun's (or murein) lipoproteins (BLPs) connect the outer membrane and the peptidoglycan layer [76, ch. 2]. While the inner (cytoplasmic) membrane of Gram-negative bacteria is a bilayer composed of phospholipids as in Gram-positive bacteria, the outer membrane consists of phospholipids (lower/inner layer) and of lipopolysaccharides (LPSs) (upper/outer layer) [76, ch. 2]. The outer membrane can be crossed by some small hydrophilic molecules through porins: These are water-filled channels which can be unspecific or contain a binding site allowing only specific molecules to pass through [35, pp. 76–77]. The area between the outer and inner membranes, where the peptidoglycan layer is located, is called the periplasm (or periplasmic space) [35, pp. 77–78]. It is the location for proteins involved in degradation processes, substrate transport, and chemotaxis response (moving to or away from a chemical gradient) [35, pp. 77–78].

Many antibiotics inhibit the process of cell wall synthesis, thus preventing the replication of cells. The most prominent drug class targeting this process are β-lactams which include penicillins, cephalosporins, monobactams, and carbapenems. Penicillin-binding proteins (PBPs) are involved in the synthesis of the peptidoglycan layer by polymerization (transglycosylation) and cross-linking (transpeptidation) of the layers; different types of PBPs can be found in bacterial cells, which can be broadly classified into high and low molecular mass PBPs [88]. β-lactams bind to PBPs, hindering them in their function [89]. Glycopeptide antibiotics have a different mode of action: they bind to peptidoglycan intermediates, and prevent transglycosylation and transpeptidation reactions [90]. Finally, some antibiotics interfere with the cytoplasmic membrane, e.g. daptomycin [91] and polymyxins [92].

DNA synthesis (replication) Bacteria reproduce by replication of their genetic material and segregation into daughter cells. The process is initiated by making DNA strands accessible for the replication complex, consisting of the bacterial initiation protein (DnaA), the helicase loader (DnaC), the replicative helicase (DnaB), and other components [85]. Two replication forks are formed, and the replication progresses bidirectionally until the terminus region is reached, i.e. when both forks meet [82]. During this process, the proceeding of the replication forks results in overwinding of the unreplicated DNA (positive supercoils) in front of the forks and intertwining (pre-catenanes) of the sister duplexes behind the forks. These need to be dissolved, which is achieved by two type-II topoisomerases: DNA gyrase and topoisomerase IV. Topoisomerase IV is also involved in separation of the two sister chromosomes at the end of the replication process [93; 94].

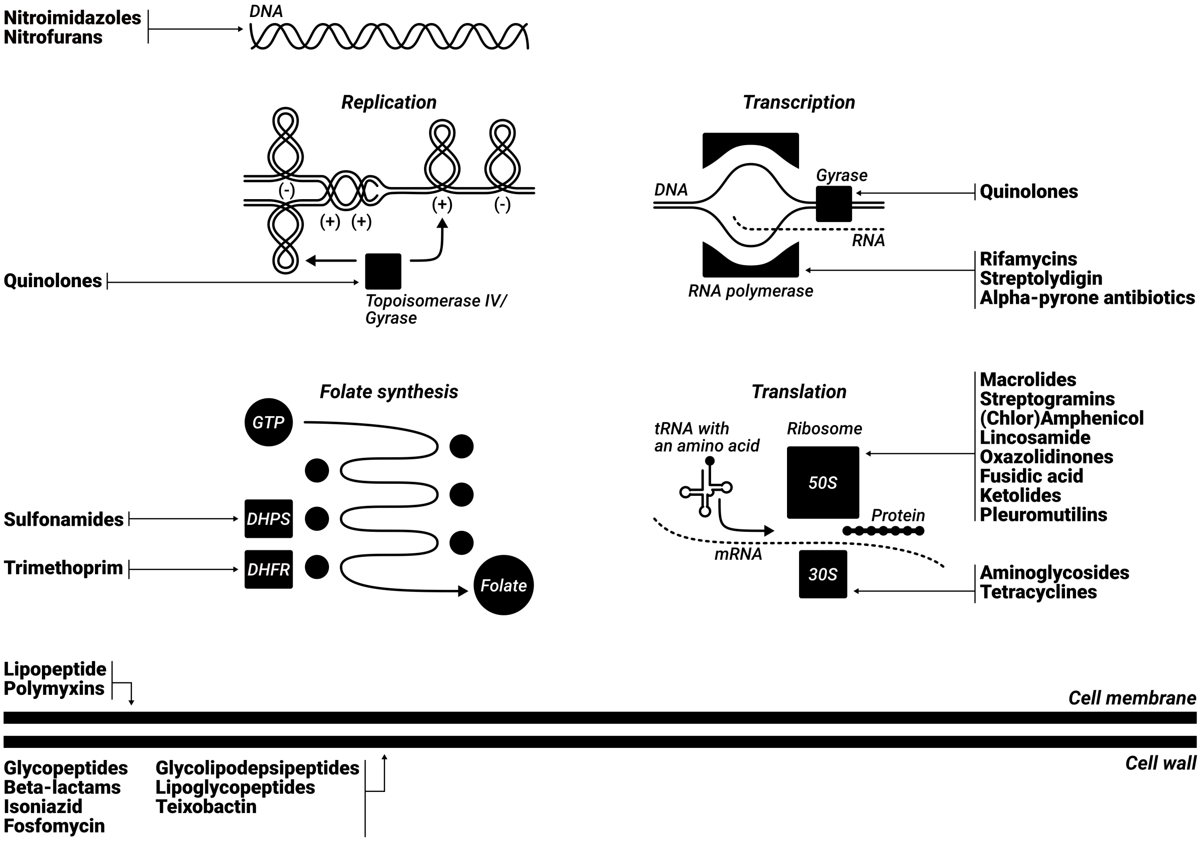


Figure 1.7: Overview of the main cell components and processes targeted by different antibiotic classes or antibiotics, adapted from [95–99].

Fluoroquinolones inhibit the replication process by binding to subunit A of DNA gyrase (GyrA) and subunit A of topoisomerase IV (ParC) [89]. Both enzymes, DNA gyrase and topoisomerase IV, act by binding to the DNA, breaking the double strand to pass through another DNA double strand, and joining the broken strands back [100]. The re-joining process is hindered by fluoroquinolones, which bind to the enzyme in the enzyme-DNA complex [100]. Depending on the bacterium and the antibiotic, either GyrA or ParC is more sensitive to the drug [89; 100; 101].

RNA synthesis (transcription) The first step for gene expression in all organisms is transcription — the synthesis of RNA from a DNA template by a RNA polymerase. The transcription is initiated through binding of the RNA polymerase to a σ initiation factor, enabling the subsequent binding to a promoter region and formation of a (closed) promoter complex [102]. The DNA double strand is then unwound upstream of the transcription start site (open promoter complex) and the template strand is guided towards the active center of the RNA polymerase [102]. The elongation complex is created through dissociation from the promoter region and the σ initiation factor [102; 103]. The transcription process is finished when a termination sequence is reached or through binding to a termination factor [102; 103]. As during replication, positive supercoils are created in front of the transcription complex [104] which are resolved by the gyrase [101].

The RNA polymerase is a target for multiple antibiotics such as rifamycins, sorangicin, streptolydigin, and myxopyronin [105]. Rifamycins and sorangicin bind to the RNA polymerase and prevent the growth of the nascent RNA molecule [105]; streptolydigin inhibits transcription initiation and elongation, and pyrophosphorolysis [106]; myxopyronin hinders transcription by preventing the formation of the initiation complex [107]. Finally, as the gyrase is also involved in the process, RNA synthesis can also be hindered by fluoroquinolones [108].

Protein synthesis (translation) During translation, the ribosome is one of the key players directly involved in protein synthesis. In bacteria, this is the 70S ribosome, which is composed of two sub-units: the 30S sub-unit is a complex of 16S rRNA and 21 proteins, and the 50S sub-unit consists of two rRNAs, 5S and 23S, and 36 proteins [62]. Prior to the translation's start, the 70S complex is split into its sub-units 30S and 50S through binding of the initiation factors IF3 and IF1 to 30S [109]. Then, IF2, mRNA, and fMet-tRNA (initiator tRNA) form the pre-initiation complex with 30S [109]. It is subsequently transformed into the 70S initiation complex by dissociation of the initiation factors and binding to the 50S sub-unit [109]. During the elongation phase, the mRNA is moved through the ribosome and the protein product is synthesized [109]. After reaching a stop codon, the translation stops releasing the created protein and the mRNA [109].

Various antibiotic classes inhibit the protein synthesis in bacteria by binding to the rRNAs in the ribosomal sub-units. The antibiotic target sites can be grouped into three main categories: antibiotics targeting the ribosomal decoding site on the 30S sub-unit which hinder codon recognition (e.g. aminoglycosides), drugs binding to the peptidyl transferase center in the 50S sub-unit which stop generation of the peptide bonds (e.g. oxazolidinones and chloramphenicol), and antibiotics interfering with the peptide exit tunnel on the 50S sub-unit which inhibit the synthesis of the translated protein (e.g. tetracyclines) [110].

Folic acid metabolism The biosynthesis of folic acid (or folate) is essential for most bacterial cells [97]. The resulting compound is used to produce some amino acids, thymidine and purine [97], and is thus essential for the synthesis of DNA, RNA and proteins and, consequently, for cell growth and division. In the folate biosynthesis pathway, guanosine triphosphate is transformed to tetrahydrofolate (tetrahydropholic acid) through multiple steps involving, among other enzymes, dihydropteroat synthase (DHPS) and dihydrofolate reductase (DHFR) [97]. These two enzymes are the targets of sulfonamides and diaminopyrimidines (e.g. trimethoprim), which disrupt the folate synthesis. While sulfonamides compete with the substrate for DHPS, diaminopyrimidines bind to DHFR [111].

1.3 *Antibiotic resistance*

Despite the discovery of antibiotics, the risk of severe bacterial infections has not been eliminated, because of emerging and spreading antibiotic resistance. The term "antibiotic resistance" describes the phenomenon of bacterial cells being non-susceptible to the applied antibiotic. It is important to note that this phenotype is not a binary but a quantitative trait which is usually expressed in the terms of the drug concentration tolerated by the organism considered. A commonly used measure for the degree of antibiotic susceptibility or resistance is the minimal inhibitory concentration (MIC) value, defined as the minimum concentration of the drug required to inhibit visible growth [112]. Depending on the number of antibiotics and antibiotic classes to which a pathogen is resistant, it can be classified as multi-drug resistant (MDR, resistant to several antimicrobials or drug classes), extensively resistant (XDR, a more extreme case of multi-drug resistance), and pan-drug resistant (PDR, resistant to all antibiotics). However, no consistent definition of these terms exists, and different variations of them can be found in the literature [113].

Some bacteria can be intrinsically resistant to an antibiotic agent. Intrinsic resistance mechanisms include the absence of the antibiotic target (e.g. by having a version of the molecule which is insensitive to the drug), efflux pumps removing the drug from the cell, and an outer membrane which is impermeable to the drug [114]. Gram-negative bacteria, for example, are intrinsically resistant to vancomycin, a glycopeptide antibiotic, because the outer membrane hinders the drug from entering the cell and reaching its target [114].

Bacteria can, however, also acquire resistance factors [115]. The development of antibiotic resistance is thus a natural process of adaption of bacteria to the environment [115]. The application of a drug creates a selective pressure on the microbial community favoring the growth and replication of bacteria less susceptible to the agent used

[115]. The next sub-sections describe the main resistance mechanisms, the plasticity of the bacterial genome allowing fast adaption to antibiotic exposure, the processes contributing to resistance emergence and dissemination, and the implications of rising levels of antibiotic resistance.

1.3.1 Resistance mechanisms

There are four main strategies by which bacterial cells can survive when exposed to an antimicrobial agent: prevention of drug accumulation, drug modification, target modification, and bypass (Figure 1.8) [115; 114].

Drug accumulation prevention A bacterial cell can protect itself from an antibiotic by becoming impermeable to the drug [115]. For example, porins (1.2.1) are often used by hydrophilic antibiotic agents to reach the interior of the bacterial cell: a change in their type, number, and efficiency can affect the ability of the drug to penetrate the outer membrane [115]. The antibiotic can also be removed from the cell by efflux systems which can pump out a specific drug (e.g. TetA efflux pumps, which can cause high-level resistance to tetracycline [116]), or a broad range of antibiotic compounds (e.g. the multi-drug resistance efflux pump MexAB-OprM in *Pseudomonas aeruginosa* [117]).

Drug modification Specific enzymes can either degrade the drug molecule or change its chemical structure hindering the drug from binding to its target [115]. For example, β-lactamases, a large group of enzymes with many classes demonstrating different substrate specificity, mediate resistance to β-lactams as they destroy the antimicrobial [118]. In the case of aminoglycosides, the drug's structure can be modified by specific enzymes through acetylation, phosphorylation, or adenylation, which decreases its affinity to the target molecule [110].

Target modification The binding between the antibiotic and its target can also be prevented by acting on the target itself. This can be achieved through "target protection", where another molecule removes the antimicrobial from the target or competes with it for the target's binding site [115]. An example of the first case are the tetracycline resistance factors Tet(M) and Tet(O), which dissolve the bond between tetracycline and its target, the ribosome [115]. The quinolone resistance protein Qnr falls into the second category — it protects the quinolone targets GyrA and ParC by binding to them [119]. Moreover, this resistance factor also facilitates the emergence of chromosomal mutations in the targets leading to higher levels of resistance [120].

A modification of the target's structure can also prevent the antibiotic from binding to it. This can be achieved through a mutation of the DNA sequence of the target, as it was observed for rifamycin, where mutations in the *rpoB* gene, which encodes a subunit of the RNA polymerase, reduced the binding affinity of the drug to the RNA polymerase [102]. The target's structure can also be altered through an enzyme. For example, the erythromycin ribosomal methylation enzyme performs a methylation of a residue in the 23S rRNA (part of the 50S sub-unit of the bacterial ribosome) resulting in resistance to macrolides, among other antibiotics [115].

Bypass The effect of an antibiotic can be avoided by increasing the expression level of the target or by producing molecules which have the same function as the target molecule, but are not affected by the drug [115]. A good example of the first case is the over-expression of DHFR (1.2.1), which is achieved through a mutation in the associated promoter [69]. Bacterial cells overproducing DHFR can achieve resistance to trimethoprim [69]. An example of the second case is the resistance to most β-lactams in *Staphylococcus aureus*, which can be gained through acquisition of the *mecA* gene which encodes PBP2a, a specific PBP. This protein has a low affinity to many β-lactams, such that these antibiotics cannot inhibit the cell wall synthesis [115].

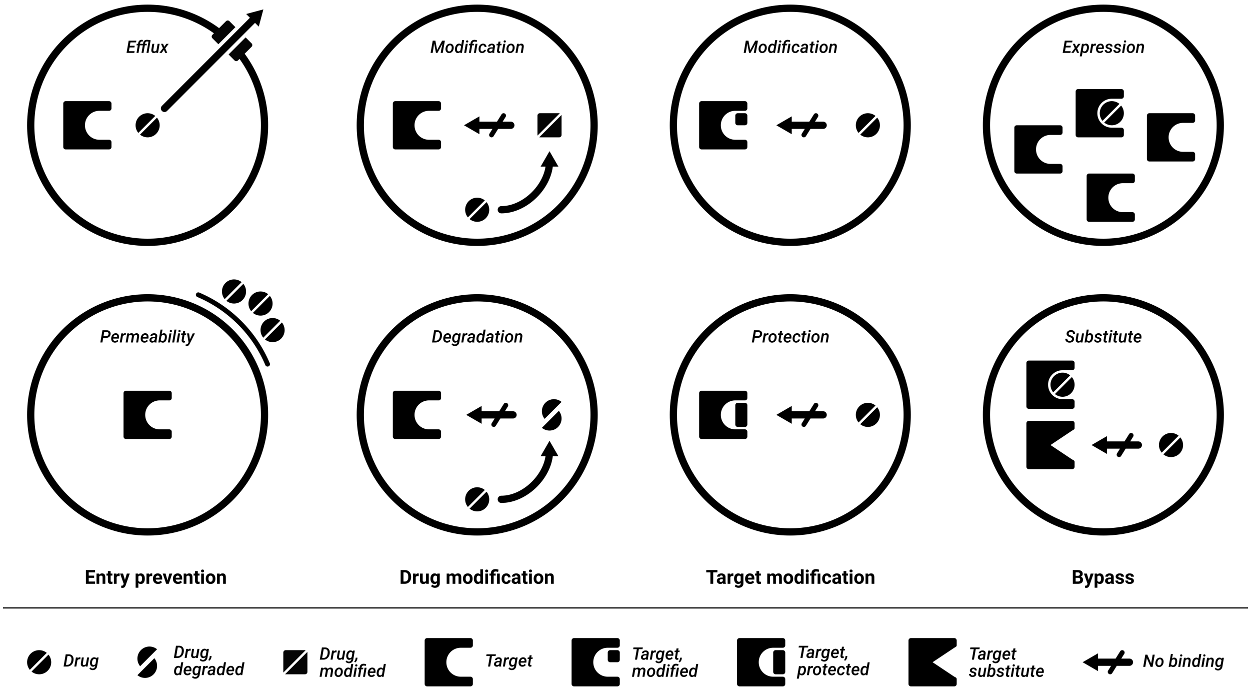


Figure 1.8: An overview of the antibiotic resistance mechanisms, adapted from [121].

1.3.2 *Bacterial genome, adaption, and resistance emergence*

As mentioned above, acquired resistance is the result of an adaption process to environmental changes. A good example of the resistance evolution over time was shown by Baym *et al.* by using large plates with a medium containing successively increasing concentrations of an antibiotic [122]. The experiment showed the emergence of new mutants during the advancement of *E. coli* cells from a drug-free region to the region with the highest drug concentration, over the course of several days [122]. The bacteria owe their remarkable adaption ability to their genome plasticity. Through mutations, genomic rearrangements, and accommodation of new genetic material, they can survive in environments with conditions which were initially lethal for them. As demonstrated by Baym *et al.* , this can also occur over a relatively short timespan [122].

Bacterial genome The size of bacterial genomes is greatly variable. Currently, from the complete set of bacterial genomes stored at the National Center for Biotechnology Information (NCBI) [1], *Candidatus Hodgkinia cicadicola* has the smallest genome, at $\sim$ 0.1 Mbp (CP025310.1), *Minicystis rosea* has the largest genome, at $\sim$ 16 Mbp (CP016211.1), and the average genome size is about 4 Mbp. GC content, the proportion of guanine/cytosine bases in the genome, varies greatly between different species, ranging from 13.5 % (*Candidatus Zinderia insecticola CARI*, CP002161.1) to 75.3 % (*Cellulomonas sp. Z28*, NZ_CP039291.1, CP039291.1). Lower GC content is also associated with smaller genome size [123].

Genes have an average sequence length of about 1 kbp [123], with a typical average density of 0.8 to 1.2 genes per 1 kbp [124]. The organization density of a genome can also be expressed in terms of the length of the intergenic regions, which vary from 0 bp and to over 1 kbp [124]. Larger regions usually contain repeats (e.g. CRISPR repeats) and pseudogenes (gene copies affected by gene decay, e.g. through mutations etc. [125]) [124].

The gene content of a bacterium is not fixed, i.e. not all genes are found in all bacteria belonging to the same species; some genes may be unique to a specific bacterial strain. This is the result of the ability of bacteria to lose, gain, and exchange genetic material. Genes can be broadly divided into two groups: core (essential) genes and accessory (dispensable) genes. The first group includes genes shared by all the organisms of a bacterial group, usually including highly conserved genes involved in essential cellular functions such as replication and growth [126]. Accessory genes include all non-ubiquitous genes which usually contribute to the adaption ability of the bacterial cell [126].

The collection of all genes observed within a bacterial group is called the pan-genome [126]. A pan-genome can be characterized in terms of whether it is closed (finite) or open (infinite) [126; 127]. Closed pan-genomes require only a minimal number of genomes to be completely described (i.e. including a new genome would not provide any new genes), while open pan-genomes can be extended by adding a new genome which increases the pool of the accessory genes [126; 127]. Bacterial species demonstrating a less pronounced specialization to a specific environment have an open pan-genome [126] as, for example, *E. coli*, which can be found both within the animal gastrointestinal tract and in non-host environments [128]. Species with a high degree of specialization (e.g. *B. anthracis*, an obligate vertebrate pathogen [76, ch. 15]) have a more conserved

[1] Data was obtained from on 27.06.2019 `https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/` using the filters "Kingdom" (set to "Bacteria") and "Assembly level" (set to "Complete").

genome and thus a closed pan-genome [126].

There are various adaption mechanisms which shape the bacterial genome, helping the organism to adjust itself to new environmental conditions.

Mutations Mutations occur continuously over the life cycle of a cell with an observed rate of adaptive mutations of about 10^{-5} events per cell generation and with a higher frequency (10^{-3} to 10^{-5} per cell generation) for mutations resulting in genomic rearrangements [129]. Since only a small portion of mutations is advantageous for the bacterium, the mutation rate is the result of an equilibrium between the negative effects of these changes and their potential benefits [130]. Under specific conditions, strains with a high mutation rate (so-called "mutators") can have a selective advantage. Mutators arise from strains with altered genes responsible for DNA repair and replication [130], such as in *dnaQ* (a DNA polymerase III subunit) [131].

Mutations directly linked to antibiotic resistance include those which result in increased gene expression or structural changes of the associated protein, affecting the interaction with the antibiotic (e.g. drug uptake or its binding to the target). If the target is encoded by multiple gene copies, mutations in only one or few of them may only result in a titration effect.

An example of mutations affecting drug uptake can be found in *P. aeruginosa*: resistance to imipenem can be achieved through a mutation in the *oprD* gene which encodes an outer membrane porin used by carbapenems to enter the bacterial cell; mutations changing the structure of this channel result in a decreased drug influx [132].

Fluoroquinolone resistance is a good example of mutations preventing drug binding. The resistance to this antibiotic class is often linked to mutations in the target site of the genes *gyrA* and *parC* [133], which have an impact on the protein structure and, therefore, also on the binding affinity to the antibiotic [100].

Genomic rearrangements Genomic rearrangements include events which remove or add sequence segments (deletions and insertions), change the copy number of a segment (duplications and amplifications), invert a sequence (inversions), and move sequences to other locations (translocations) [134]. These modifications can affect the chromosome structure and the expression rate, and function and copy number of a gene; they can also disrupt gene sequences and add a new sequence to the genome [134]. The sequence segment involved in a rearrangement event can vary in length and contain, when spanning an intergenic region, a few or many genes and even complete operons (clusters of genes controlled together) [135].

HGT Horizontal gene transfer (HGT) is a process whereby the genetic material is not transferred vertically, i.e. from the parent to an offspring, but through a lateral transfer [136]. For bacteria, there are several ways in which this process can be mediated: DNA can be absorbed from the cell's environment (transformation), transported from one cell to another by bacteriophages (transduction), exchanged between bacterial cells connected by a pilus (conjugation) or through nanotubes, transferred by specific gene transfer agents, and collected by the uptake of membrane vesicles carrying genomic sequences [134].

A successful HGT event requires integration of the new DNA segment into the recipient bacterial genome [137]. Moreover, to withstand the selective pressure the acquired gene needs to be expressed and, optimally, provide its host with an adaption advantage [134]. Otherwise, the new DNA sequence can again be removed from the genome [134; 137].

HGT generally takes place between cells which co-exist in the same environment. It can even transfer genetic material between distantly related organisms, though it is less frequent than between closely related bacteria, as a dissimilarity in sequence composition may hinder or hamper expression of the new genes [134]. Overall, acquisition and utilization of foreign DNA is limited by factors such as the ability of the host bacterium to accommodate DNA from the environment or from other bacteria, possible degradation by the host's restriction enzymes (endonucleases), and successful maintenance (replication) and expression [138].

Mobile elements Mobile elements contribute to the genomic rearrangement events and HGT. As their names suggests, they move DNA segments either within or between genomes, can undergo recombination events with each other, and encompass a variety of different elements [139; 134], as described below. [2]

Insertion sequences (ISs) are DNA sequences containing a transposase gene (which mediates the mobility of the IS [140]) flanked by short terminal inverted repeats. Their insertion can affect the expression (positively or negatively) of a gene or neighboring genes, and lead to genomic rearrangements; their inexact removal can result in insertions or deletions.

Miniature inverted-repeat transposable elements are small (up to 0.5 kbp) and AT-rich DNA sequences with short terminal inverted repeats. They can have a similar effect on the genome as the insertion sequences.

Repetitive extragenic palindromic sequences (REPs) are sequences of palindrom (forward sequence is the same as its reverse complement) repeats. These palindroms are imperfect and have a length of 20 b to 40 b. REPs frequently occur in pairs or clusters, and a pair of REPs found in inverse order with a linker sequence between them is also called a bacterial interspersed mosaic element. The distribution of the REPs and bacterial interspersed mosaic elements is organism specific, and these elements are assumed to be involved in evolutionary processes and affect genome stability.

Transposons represent the DNA sequences of several kilobases (usually between 2.5 kbp and 60 kbp) containing multiple genes flanked

[2] The following summary of mobile elements (up to the end of this paragraph) is based on an extensive review of Darmon and Leach [134]. Thus, this reference is omitted in the subsequent text passages and only other additional references are listed.

by long terminal inverted repeats. The changes induced by transposons are analogous to those of ISs, but they also can mediate the insertion of new genes into the host's chromosome.

Transposable bacteriophages are viruses which infect bacterial cells and integrate their DNA into the genome of the host. They can either enter a lysogenic cycle and remain in the host's genome, or start producing new phages which ends in the lysis of the host cell (lytic cycle).

Genomic islands (GIs) are DNA sequences spanning between 10 kbp and 200 kbp, which are found in some bacterial strains but not in other closely-related strains. These DNA segments are usually located on the bacterial chromosome, and demonstrate different sequence-based characteristics (e.g. codon usage and GC content) to the rest of the genome, which indicates their "foreign" origin. GIs carry multiple accessory genes which are beneficial to their hosts.

There are also other genomic sequences which can, according to Darmon and Leach, be classified as mobile elements, including inteins and retroelements. An additional and important group of mobile elements are plasmids, which are described subsequently.

Plasmids New genetic material can also be acquired in form of plasmids, DNA molecules which can replicate autonomously (replicons) [139; 134]. Plasmids are generally circular and smaller than bacterial chromosomes [139], and usually contain some conserved genes (the "backbone") and accessory genes [141]. The latter often encompass genes which might be useful to the host, including virulence and antibiotic resistance factors, while the "backbone" genes are linked to the plasmid's replication and mobility mechanisms [139; 141]. Plasmids sharing similar replication and segregation mechanisms cannot co-exist in the same cell, as they interfere with each other during cell division. In this case, they are labeled incompatible [142]. A plasmid can also posses multiple replication mechanisms, enabling it to replicate across a broader range of different bacterial hosts [142]. Plasmids are usually exchanged between bacterial cells through conjugation, though transformation is also an option [138]. Depending on whether they posses the genes required for this process, plasmids can be classified as conjugative (all genes needed are present), mobilizable (other conjugative elements are needed for transfer) or non-mobilizable (not able to transfer through conjugation) [142; 143]. The transmission frequency of plasmids can be influenced by multiple factors leading to higher or lower dissemination rates [142]. These factors include co-integration (recombination of co-existing plasmids), which can also enable the transmission of non-mobilizable plasmids [142].

The recent discovery of the plasmid-mediated gene *mcr-1* illustrates not only the spread of a relevant resistance factor through plasmids, but also the emergence of a new resistance determinant. The gene was first described by Liu *et al.* in 2015, who found it in Enterobacteriaceae isolated from livestock animals [144]. It confers resistance to colistin, a polymyxin, considered to be one of the "last resort" antibiotics

[144]. Previously, only chromosome-located resistance mechanisms had been reported [144]. After the discovery of *mcr-1*, the gene was detected in different Enterobacteriaceae species worldwide [145], and other studies reported new variants of this gene [146–153].

It has been hypothesized that the main origin of resistance determinants are environmental bacteria which harbor many of these factors on their chromosomes [154]. The more specialized pathogenic bacteria were initially susceptible to antibiotic exposure [154]. However, the extensive production and use of antibiotics has promoted the acquisition of resistance factors by pathogens, also favoring their dissemination and exchange [154].

1.3.3 Implications of growing antibiotic resistance

Rising levels of antibiotic resistance lead to less effective or failing treatment for bacterial infections, and have become a serious threat worldwide [155]. The major pathogens which cause most nosocomial infections are *Enterococcus faecium*, *S. aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *P. aeruginosa*, and *Enterobacter* species — abbreviated to ESKAPE pathogens [156]. These and some additional pathogens (*Clostridium difficile*, *Neisseria gonorrhoeae*, *Campylobacter*, *Salmonella*, *Shigella*, *Streptococcus pneumoniae*, *Mycobacterium tuberculosis*) were considered as "threatening" (ranked from "urgent" to "concerning") to the US Centers for Disease Control and Prevention (CDC) in 2013 [157]. A similar list of pathogens was published in 2017 by the World Health Organization (WHO), highlighting bacteria and the associated resistance phenotype which would require the development of new antibiotics [158].

According to a recent report by Jason *et al.* , infections caused by multi-drug resistant organisms were the third highest cause of deaths in the US in 2010 [159]. A review on antimicrobial resistance commissioned by the UK government and published in December 2014 estimates 300 million deaths due to resistant pathogens and a substantial negative impact on the world's GDP between 2014 and 2050 [160]. Moreover, studies which have been considered in this review and model the increase in antimicrobial resistance may underestimate the true costs, because of a lack of available data about (bacterial) infections [160]. For example, the RAND report included focused only on hospital-acquired infections caused by *E. coli*, *K. pneumoniae* and *S. aureus*, and drug resistance related to HIV, tuberculosis and malaria [161]. Other implications described in the review consider the overall effect of the increasing frequency of resistant pathogens on medical procedures which rely heavily on antibiotics for prophylaxis [160]. Interventions with a high infection probability (e.g. surgery) or those suppressing the immune system (e.g. some cancer therapies) would become much more dangerous for the patient [160]. In summary, the review urges taking action in order to tackle the resistance problem and limit its negative consequences [160].

1.4 Fighting bacterial resistance

The problem with antibiotic resistance is multifactorial, including antibiotics misuse (prescriptions for non-bacterial infections) and overuse [162], the extensive use of antibiotics in agriculture and livestock farming [162], factors affecting infection transmission such as population density, sanitation (access to clean water, sewage systems etc.) and travel [163], and a stagnation in the development of new antibiotics [164].

There are a number of strategies to counteract the implications of antibiotic resistance. Since there are multiple aspects contributing to the problem it has to be looked at from different angles, including preventive measures to reduce the number infections, surveillance projects to observe resistance trends, diagnosis and therapy development to increase treatment effectiveness, and research projects studying bacterial pathogens.

1.4.1 Vaccination

Vaccination targeting the relevant bacterial pathogens, such as *Mycobacterium tuberculosis*, *S. aureus* and *A. baumannii* [165], would reduce the number infections and antibiotic use, thus lowering selective pressure promoting the emergence of resistance. Currently, vaccines for *Haemophilus influenzae* [166], *Salmonella typhi* [167], and *S. pneumoniae* [168] are already available. Additionally, vaccines against other infectious diseases can contribute to a decrease in antibiotic misuse [169]. As antibiotics are often prescribed for conditions not caused by bacterial pathogens, e.g. influenza, preventing these infections could help in minimizing unnecessary antibiotic treatments [169].

1.4.2 Therapy development and optimization

Effective therapies are one of the key components in combating infectious diseases. Besides the development of new antibiotics, there are also other approaches such as therapy optimization and alternative treatment strategies using bacteriophages, antimicrobial peptides and monoclonal antibodies.

New antibiotics The discovery and development of new antibiotic agents would extend the palette of available drugs, providing more options for the effective treatment of bacterial infections. As mentioned previously, however, the antibiotic development has stagnated. Candidates for new drugs are mostly derived from the already discovered classes [164] and the few new classes, which were launched between 2000 and 2012, were only for Gram-positive bacteria [170]. Moreover, it has been argued that the number of groups working at pharmaceutical companies and looking for new antibiotic drugs has decreased due to mergers of these companies [41]. In addition to the shortage of new compounds, the complete pipeline of drug discovery, optimization, testing and approval requires several years and major investments [41]. As a result, some companies have backed out of antibiotic discovery and focused on medication required for chronic diseases (e.g. high blood pressure), this being a much more stable source of income [41]. Initiatives have been launched to encourage and facilitate the development of new antibiotics such as, for example, the GAIN ("Generating Antibiotic Incentives Now") project started in the United States in 2012 and ND4BB ("new drugs for bad bugs"), started by the European Innovative Medicines Initiative in the same year [41].

Spaulding *et al.* argue that there is also a need for "precision antimicrobials" [21]. The application of broad-spectrum antibiotics has an adverse impact on a patient's health: severe changes in the host's microbiota can lead to opportunistic infections and contribute to antibiotic resistance development [21]. These problems should be avoided through a more targeted approach, which would eliminate specific pathogens without affecting other microbes [21]. However, the development of these antibiotics faces even greater challenges than in the general case, as it defines strict constraints on the expected properties of the active agent [21].

Drug combinations When considering the mode of action for the antibiotics and the known resistance factors, the observation was made that antibiotics targeting only a single protein are less successful than those interacting with multiple or complex targets [171; 89]. A resistance can manifest much more easily if a single change in the target is sufficient to prevent the antibiotic from interfering with it; modifications of multiple targets or those affecting all the gene copies of a target are less likely to occur [171]. While this favors the use and development of multi-target drugs, another implication is that combining multiple antimicrobials can also improve treatment effectiveness. The administration of antibiotics has generally been a monotherapy, and using a combination of drugs could extend the application spectrum and prevent resistance emergence [171]. Tyers and Wright describe in their review three classes of combinations with improved antibacterial effect: congruous combinations where the individual active agents have an antimicrobial effect (e.g. combination of penicillin and streptomycin), syncretic combinations where one of the compounds is not an antibiotic (e.g. compounds inhibiting β-lactamases), and coalistic combinations where none of the compounds is an antibiotic but their combination has an antimicrobial effect [171]. The development of effective drug combinations is challenging, with increasing complexity for higher-order combinations, though it is an important strategy when considering the rising levels of resistance [171].

Other approaches Antibiotics represent only one possible way of fighting bacterial infections, and there are many other approaches which could be applied alone or combined with antibiotic treatment.

In phage therapy, bacteriophages (see 1.3.2) are used to treat bacterial infections [172]. Though some of these viruses can enhance the virulence of the bacteria [173], there are also phages which kill their host by entering a lytic life cycle, leading to the lysis of the bacterial cell [172]. The virus replicates by exploiting the host: after a successful infection, replication processes are induced, the host's cell membrane is dissolved, and the newly-created phages are released into the environment [174]. The main advantages of this form of therapy are the amplification of the active agent through the cells which it should eliminate, and the host specificity of some phages [174]. However, there are also challenges and drawbacks, such as the need to account for emerging phage-resistance, the requirement of a phage susceptibility test to ensure effective treatment, the risk that bacterial cell lysis can also release harmful substances, and the problem that not all infections can be treated with phage therapy (e.g. those where bacteria infect eukaryotic cells which cannot be entered by phages) [174].

Antimicrobial peptides are naturally-occurring molecules of 15 to 50 amino acids which can be found in many species and are involved in host defense against pathogens [175; 176]. These natural antimicrobials can have a direct effect on the pathogens or an indirect effect by hindering the formation of biofilms or destroying them [177]. A special group of antimicrobial peptides are bacteriocins, antimicrobials produced by some bacteria [178], though others argue that they are "full-blown" proteins rather than peptides [175].

Another alternative approach is targeted therapy with monoclonal antibodies which were used before the discovery of antibiotics ("serum therapy") but have been replaced by them, due to their ease of production and administration [179]. Potential targets of the antibodies are molecules released by bacterial cells (e.g. toxins) and cell surface components [180]

Recently, Ragheb *et al.* proposed the treatment strategy of "inhibiting evolution" [181]. For many drugs, antibiotic resistance is caused by *de novo* mutations which are facilitated through "evolvability factors" [181]. Inhibiting these would negatively affect the adaptability of bacteria, preventing the emergence of resistance associated mutations [181]. When administrated together with an antibiotic, evolution inhibitors could improve the effectiveness of antimicrobial therapies [181].

1.4.3 Diagnosis: pathogen identification

In diagnosis, the identification of the pathogen responsible for the infection is vital in order to select an optimal therapy. It starts by differentiating between bacterial and viral infections to avoid unnecessary use of antibiotics. Further characterization of the bacterial pathogen, such as its strain type and the resistance factors contained in its genome, can help to determine which antibiotics are likely to be ineffective against it, thus reducing their use.

The classical pathogen identification procedure involves sample collection, organism isolation and culturing, followed by identification [182]. Automated systems which facilitate diagnosis have been developed such as Vitek 2 (bioMérieux, Marcy l'Etoile, France) and Phoenix (BD Diagnostics, San Jose, CA, USA) [183]. MALDI-TOF mass spectrometry (MS) systems have been proposed and applied, including the Bruker Microflex Biotyper (Bruker Corporation, Billerica, MA, USA) and the bioMérieux VITEK MS [184].

As culturing is a time-consuming process, culture-independent approaches have been developed. These detect specific DNA regions (usually amplified by PCR) or antigens [185]. Culture-independent tests have multiple advantages compared to culture-based tests: they decrease the required costs and time, are easier to apply (thus promoting their use), can have higher sensitivity, are applicable if there is a pathogen with no established culturing protocol, and have the ability to detect multiple organisms (multiplex tests) [186; 187; 185]. However, there is also a major drawback: these tests do not provide further characteristics of the pathogens which are relevant for surveillance projects [187; 186].

When a more detailed characterization of the bacteria is required, e.g. information on the strain types for epidemiological studies, genotyping becomes necessary, which can be done using non-sequence-based and sequence-based methods [188]. However, only whole-genome sequencing (WGS) can provide the complete genetic information of an isolate (see 1.4.5). It allows for taxonomic characterization, exhaustive sub-typing of the samples, detection of relevant genetic factors such as virulence and resistance genes, and comparison of the samples to infer possible transmission of the pathogens during an outbreak [188].

The value gained from performing WGS comes at the cost of depending, again, on the culturing step. Culture-independent metagenomic analysis can be used to obtain genomic data of microbes found in a sample while bypassing culturing and not limiting focus to a specific organism [186]. Though this approach is promising and has already been applied in clinical setting [189–191], its application remains challenging. Problems include host contamination and sample-specific factors, such as the sample type (e.g. urine, blood) which affects the processing procedure [186], and high variance in community composition even among healthy individuals [192]. There are also further challenges regarding the analysis of sequencing data, which are discussed in more detail in 1.4.5.

1.4.4 Antibiotic resistance surveillance and stewardship

The implementation of effective preventative and counteractive measures requires an exhaustive knowledge of bacterial resistance mechanisms, and information on their emergence, spread and prevalence. The collection and analysis of the associated information is therefore crucial to limiting and combating the implications of antibiotic resistance. Systematic monitoring of bacterial infections allows for the recognition of trends of rising resistance rates and the prevalence of specific pathogens. Worldwide, healthcare agencies have started surveillance programs to collect data about (nosocomial) infections and antibiotic use. There are surveillance projects on antimicrobial resistance and antimicrobial consumption in Europe by the European Centre for Disease Prevention and Control (ECDC) [193], and similar initiatives are being conducted by CDC in the United States [157]; in 2015, the WHO launched the GLASS (Global Antimicrobial Resistance Surveillance System) project to track resistant pathogens in different countries [194]. More extensive action plans have also been initiated, such as the One Health Action Plan against antimicrobial resistance in the EU [195] and the "National action plan for combating antibiotic-resistant bacteria" in the US [196].

1.4.5 Study of bacterial isolates and communities

As detailed over the previous sub-sections, knowledge about resistance mechanisms and other characteristics of bacteria is vital to prevent and cure infectious diseases. Therefore, studies of specific bacterial species, especially human pathogens, and complete bacterial communities should be seen as one of the essential steps towards a solution to the antibiotic resistance crisis.

Bacterial isolates Analysis of bacterial isolates has been widely applied for association studies [197; 198] and follow-up studies of outbreaks [199; 200]. A common approach is to perform WGS of the isolates, followed by a reconstruction and analysis of their genomes. Short-read sequencing has often been performed using Illumina instruments (e.g. MiSeq and NextSeq), producing reads with a maximal length of 300 bp and an accuracy of around 99.9 % [201]. As short reads cannot span repeats, the reconstructed genome assemblies often remain fragmented [201]. Long-read sequencing with lower accuracy but much longer reads allows for the creation of complete, i.e. full-length, genome assemblies and can be done using the SMRT platform from PacBio and the MinION from Oxford Nanopore Technology's [201]. Also, a hybrid approach, which combines the data produced by the Illumina and MinION platforms, has been applied to resolve the structure of bacterial genomes [201].

Given the sequencing data, it is possible to use reference-based, reference-guided, and reference-free methodologies. If the sample is known to be closely related to a specific reference genome and only minor discrepancies are expected between them, the sequencing reads can be directly mapped to that reference genome. In other cases, if no suitable reference genome is available or the choice of a reference genome is not straightforward, reference-free approaches are more appropriate. *De novo* assembly of the genomes is then usually performed, followed by assembly annotation including the identification of protein coding sequences. Downstream processing of the collected data can include comparison to other isolates, taxonomic characterization and strain typing (e.g. multi-locus sequence typing (MLST)), and detecting the presence of specific genomic features (e.g. virulence genes). Additional data, such as the phenotype or sample metadata (e.g. collection date and location), can be linked with the genomic information for further analysis.

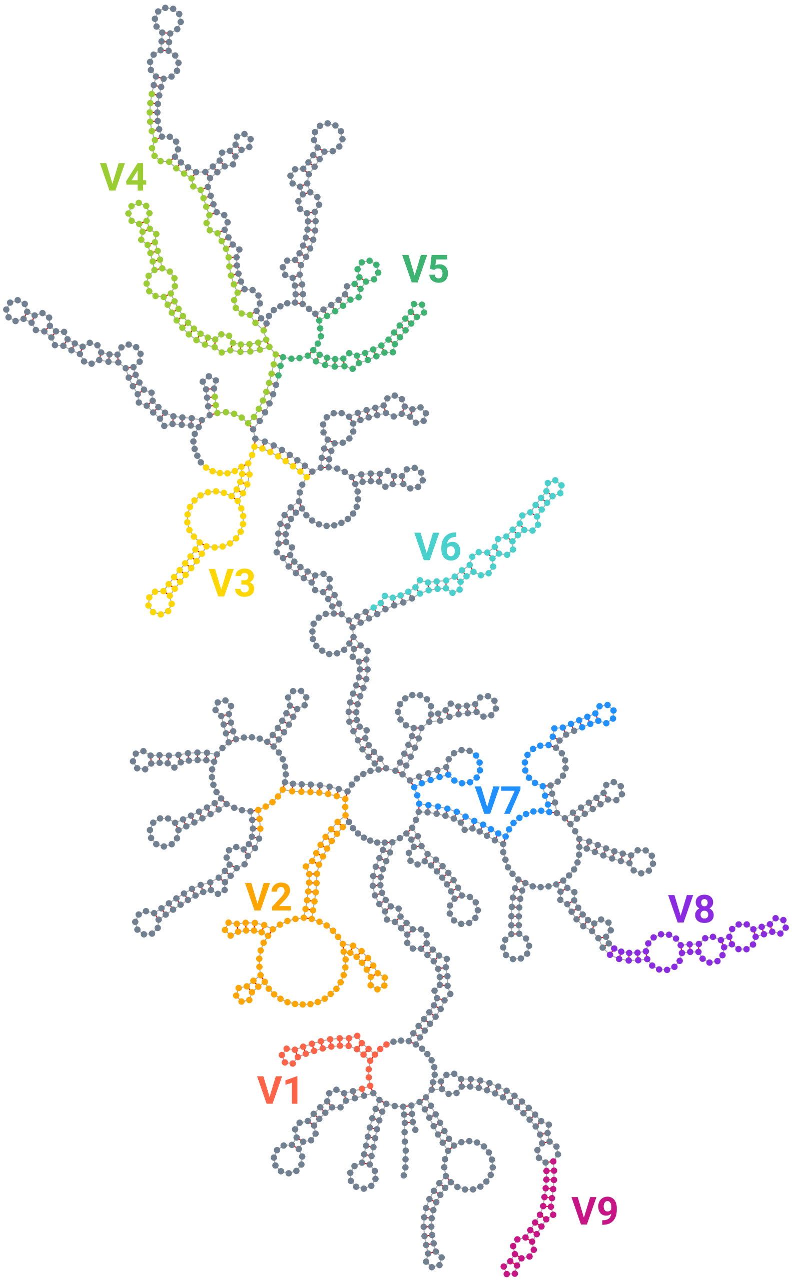
In the context of antibiotic resistance, WGS data is linked to resistance profiles of the isolates to one or multiple drugs. Profiles are determined by antibiotic susceptibility testing (AST). Classical AST methods are phenotype-based approaches, where bacterial cultures are exposed to an antibiotic visually detecting their growth [202]. The conventional methods include disk diffusion, manual broth microdilution, gradient tests, agar dilution and breakpoint tests [202]. A commonly-used measure of resistance is the MIC value, which is then interpreted using guidelines in which thresholds for susceptible and resistant phenotypes are defined for relevant bacterial groups and drugs [202]. The analysis of the bacterial genomes coupled with the associated resistance profiles can help to identify genomic features which correlate to the phenotype. Moreover, predictive models can be build to infer resistance profiles for new samples. This can facilitate the development of pathogen identification test arrays by combining relevant genetic features to be detected in the samples.

In addition to dedicated projects for performing such analyses, initiatives for collecting whole genome data and the associated resistance profiles for public use create a solid basis for resistance association studies. For example, NCBI allows for the submission of resistance profiles linked to a biological sample, and thus provides the possibility of compiling datasets containing genomic and resistance data. There also also other platforms dedicated to providing access to bacterial genomes, such as the PATRIC database which collects and processes genomes and metadata from GenBank, RefSeq and collaborators [203].

Bacterial communities The microbial communities found on and in our body play a crucial part in our health. Their composition has been linked to multiple diseases such as type 2 diabetes, obesity and colorectal cancer [204]. The microbiota colonizing our environment, including soil and water, can also be relevant with regard to bacterial infections, as these communities may function as a reservoir of antibiotic resistance genes (1.3.2). The characterization of microbiota is a challenging task as the communities are usually very complex, with hundreds of species in highly variable abundance [205]. For example, more than 100 microbial species can be found in one milliliter of water, more than 400 in the human gut, and thousands in just one gram of soil [206].

Many microbiota studies are currently carried out by sequencing specific sub-regions (variable regions) of the 16S rRNA gene [210]. The 16S rRNA gene is present in almost all bacteria [211; 212] as part of the bacterial ribosome (1.2.1) and contains conserved and variable domains (Figure 1.9); the latter being used to differentiate between

Figure 1.9: Secondary structure of the 16S rRNA, from *E. coli* with highlighted variable regions: V1 (69–99), V2 (137–242), V3 (433–497), V4 (576–682), V5 (822–879), V6 (986–1041), V7 (1117–1173), V8 (1243–1294), and V9 (1435–1465) [207]. The secondary structure was predicted for the 16S rRNA gene *rrsH* (NCBI Gene ID 944897) by RNAfold web server (version 2.4.11) using the minimum free energy algorithm [208]. The secondary structure was visualized using the Forna web server [209]. Sequence numbering and nucleotide labels were omitted in the interests of simplicity.



microbes [212]. Specific primers are designed to target and amplify a variable region prior to next-generation sequencing (NGS) [213]. Operational taxonomic units (OTUs) are subsequently defined by clustering the sequences according to their similarity, e.g. using a minimum threshold of 97 % [214]. However, this approach also has a major drawback, namely an inherently limited taxonomic resolution, i.e. it does not allow for any differentiation between (genera and) species [215; 216]. A possible solution would be to perform full-length 16S rRNA gene sequencing, which can be achieved by the use of the new third-generation sequencing technologies, such as the platforms offered by PacBio [217] and Oxford Nanopore Technologies [218]. The use of the 16S rRNA gene has also another drawback: the number of *rrn* operons (rRNA gene clusters) varies between organisms [206] which can create a bias during the sequence amplification step towards organisms harboring many copies [219].

As specific phenotypic traits can be encoded by genes belonging to the accessory and not to the core genome, capturing the complete genomic information contained in a sample has a clear advantage over 16S rRNA sequencing. Therefore, the metagenomic analysis of microbial communities has gained popularity over the last years [220]. Nevertheless, 16S rRNA sequencing has not yet been fully replaced by WGS, as it has its own challenges. One of these is community composition reconstruction, also referred to as "binning" [3]. The objective of the procedure is the assignment of sequences (e.g. long reads or contigs) to different clusters (bins), which should represent organisms or groups of closely-related organisms [3]. Current binning approaches generally differ in terms of whether they rely on a reference database (supervised) or are reference-independent (unsupervised), and in the type of features used, such as the nucleotide composition (usually five-mer frequencies) and coverage information [3; 221–223]. The quality of the bins can be assessed by estimating their purity (whether the cluster contains one or multiple organisms) and completeness (whether the cluster contains all the genomic sequences of an organism) [224].

Microbiome approaches are considered to be a new promising diagnostic tool in clinical microbiology [225; 226]. They enable the identification of pathogens in patient samples, with higher sensitivity than the currently applied methods. For example, Sathiananthamoorthy *et al.* demonstrated the benefits of 16S rRNA gene sequencing when compared to routine midstream urine culture analysis, as performed in the UK, for the diagnosis of urinary tract infections [215]. In a retrospective study, Loman *et al.* showed that high-throughput sequencing can be applied to identifying outbreak pathogens [227]. Moreover, metagenomic approaches can be used for antibiotic resistance surveillance by studying the microbiota composition of relevant environments, such as urban sewage and hospitals [228].

Bacterial plasmids In the context of antibiotic resistance, the analysis of plasmids found in bacterial isolates or communities is of particular interest. It can help to identify which resistance and virulence factors might be transferred to other cells using these mobile elements. For plasmids known to harbor highly-relevant resistance and virulence features, it is crucial to know their transmission rate, and whether they can be exchanged between relevant (human) pathogens and persist in the absence of selection pressure [144].

Finding plasmids in the WGS data of bacterial isolates or microbiota is a challenging task. The classification of genomic sequences as chromosomal or plasmid-derived can be facilitated by using resources of known plasmids. However, this can be insufficient for the identification of novel plasmids. Many existing approaches for chromosome/plasmid classification rely on differences in sequence composition and coverage between bacterial chromosomes and plasmids [229–232]. Nevertheless, there is still no gold-standard approach for the reliable detection of plasmid-borne sequences in WGS data.

Given a plasmid sample, PCR-based replicon typing has been commonly applied for plasmid characterization [233]; an alternative approach is MOB typing targeting not the replicons but relaxases [141]. The analysis can be also performed *in silico* on plasmid sequences, obtained either through direct sequencing of the plasmid or from a WGS sample [234; 235]. As with all marker-based approaches, these typing methods have the disadvantage of not considering the other genetic content found on the plasmid. Moreover, not all plasmids are type-able because the marker may be novel and not included in the resource used, or because it is absent. A more precise and specific identification can be done by using the resources of all the known bacterial plasmids, not relying on one specific genomic feature. Such resources would also have the advantage of being able to store other information related to the plasmids, e.g. collection location and host.

2
Goals of the PhD thesis

As outlined in the previous chapter, antibiotic resistance is a multifaceted problem and there are several key components which can help in addressing it. This thesis focused on the points outlined in 1.4.5: the study of bacterial isolates, microbial communities, and plasmids. The two main objectives were the analysis of a large collection of bacterial isolates and their antibiotic resistance profiles, and the development of resources enabling further studies of bacterial pathogens (Figure 2.1).

The herein analyzed sample collection contains about eleven thousand bacterial clinical isolates, including their WGS data, resistance profiles, meta-information, and, for a subset of isolates, additional taxonomic characterization by MALDI-TOF MS. The studies were performed in collaboration with Siemens AG (who collected the isolates), Curetis GmbH (who acquired the dataset), and Ares Genetics GmbH, a Curetis Group Company. This extensive dataset was used to perform various analyses, answering questions related to bacterial pathogens and antibiotic resistance: the evaluation of different WGS-based tools for taxonomic characterization of bacterial isolates and a comparison with the MS-based results (see 3.1) [1], the assessment of resistance rates with respect to different bacterial species and antibiotics, construction and analysis of bacterial pan-genomes, and linking genomic data and resistance profiles (see 3.2) [2]. A web server, GEAR-base, was implemented to provide access to the isolates collection and generated results, with the aim of extending this resource with new data by collaborating with hospitals and industrial partners (see 3.2) [2].

The other projects presented herein are not directly related to GEAR-base, and include the following three publications: a web server for binning metagenomic sequences using five-mer frequency profiles (BusyBee Web, in 3.3) [3], a review of different tools for plasmid prediction from WGS data of bacterial isolates (tools evaluation was performed on a subset of samples from GEAR-base, in 3.4) [4], and a resource of known complete bacterial plasmids (PLSDB, in 3.5) [5].

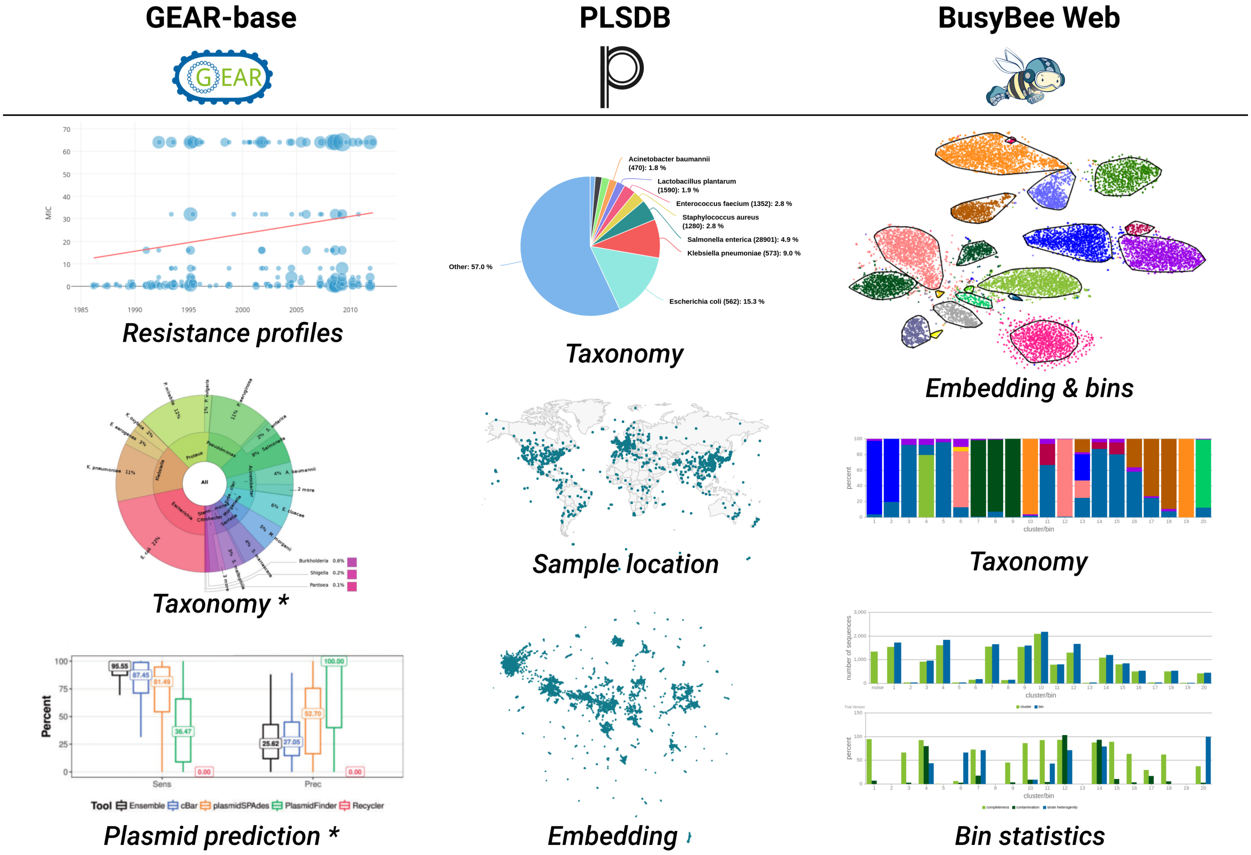


Figure 2.1: Overview of the implemented resources, including some of their functions and/or additional studies performed using their data (marked with an asterisk (*)).

3 Results

This cumulative thesis is based on five peer-reviewed publications whose published versions are included in this chapter.

3.1 *Comparing genome versus proteome-based identification of clinical bacterial isolates*

This is a pre-copyedited, author-produced version of an article accepted for publication in *Briefings in Bioinformatics* following peer review. The version of record *Briefings in Bioinformatics, Volume 19, Issue 3, May 2018, Pages 495–505* is available online at: https://doi.org/10.1093/bib/bbw122.

Comparing genome versus proteome-based identification of clinical bacterial isolates

Valentina Galata, Christina Backes, Cédric Christian Laczny, Georg Hemmrich-Stanisak, Howard Li, Laura Smoot, Andreas Emanuel Posch, Susanne Schmolke, Markus Bischoff, Lutz von Müller, Achim Plum, Andre Franke and Andreas Keller

Corresponding author. Andreas Keller, Saarland University, Building E2.1, 66123 Saarbrücken, Germany. Tel.: +49 (174) 1684638; Fax: +49 (0)6841-162-6185; E-mail: ack@bioinf.uni-sb.de

Valentina Galata is PhD student at the Chair of Clinical Bioinformatics at Saarland University.
Christina Backes is Postdoc at the Chair for Clinical Bioinformatics at Saarland University.
Cédric Christian Laczny is Postdoc at the Chair for Clinical Bioinformatics at Saarland University.
Georg Hemmrich-Stanisak is a research scientist at the ICMB at the Christian-Albrechts-University of Kiel in Germany.
Howard Li, PhD, worked as a system development technical lead at Roche Molecular Systems. His expertise includes sample preparation, laboratory automation and medical device R&D. He is a member of American Chemical Society and American Society for Microbiology.
Laura Smoot, PhD, is a microbiologist with training and research in molecular microbiology field, and has experience in in vitro diagnostics industry. She worked at Siemens Healthcare, R&D. She is a member of American Society for Microbiology (ASM) and European Society of Clinical Microbiology, and Infectious Diseases (ESCMID).
Dr Andreas Emanuel Posch is Senior Key Expert for Bioinformatics and Systems Biology at Siemens Healthcare, In Vitro Diagnostics and Bioscience R&D.
Dr Susanne Schmolke is Senior Project Manager Strategy at Siemens Healthcare. Her expertise includes virology, molecular biology and medical device industry.
Markus Bischoff is professor and senior scientist at the Institute of Medical Microbiology and Hygiene at Saarland University.
Lutz von Müller was vice head of the Institute of Medical Microbiology and Hygiene at the Saarland University. Current position: Head of Institute of Laboratory Medicine, Microbiology and Hygiene, Christophorus Hospitals, Coesfeld, Germany.
Dr Achim Plum is Managing Director of Curetis GmbH and molecular geneticist by training. His areas of expertise include precision medicine and companion diagnostics, biomarker discovery and validation, IVD industry and molecular diagnostics.
Andre Franke, PhD, is the director of the ICMB at the Christian-Albrechts-University of Kiel in Germany. The primary foci of his research are high-throughput analyses, laboratory automation, next generation sequencing, chronic inflammatory diseases, GWAS, and bioinformatics. He is a member of the German Society for Human Genetics (GfH) and the German Society for Internal Medicine (DGIM).
Andreas Keller is professor and head of the Chair for Clinical Bioinformatics at Saarland University.

Abstract

Whole-genome sequencing (WGS) is gaining importance in the analysis of bacterial cultures derived from patients with infectious diseases. Existing computational tools for WGS-based identification have, however, been evaluated on previously defined data relying thereby unwarily on the available taxonomic information.

Here, we newly sequenced 846 clinical gram-negative bacterial isolates representing multiple distinct genera and compared the performance of five tools (CLARK, Kaiju, Kraken, DIAMOND/MEGAN and TUIT). To establish a faithful 'gold standard', the expert-driven taxonomy was compared with identifications based on matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) analysis. Additionally, the tools were also evaluated using a data set of 200 *Staphylococcus aureus* isolates.

CLARK and Kraken (with *k*=31) performed best with 626 (100%) and 193 (99.5%) correct species classifications for the gram-negative and *S. aureus* isolates, respectively. Moreover, CLARK and Kraken demonstrated highest mean F-measure values (85.5/87.9% and 94.4/94.7% for the two data sets, respectively) in comparison with DIAMOND/MEGAN (71 and 85.3%), Kaiju (41.8 and 18.9%) and TUIT (34.5 and 86.5%). Finally, CLARK, Kaiju and Kraken outperformed the other tools by a factor of 30 to 170 fold in terms of runtime.

We conclude that the application of nucleotide-based tools using k-mers—e.g. CLARK or Kraken—allows for accurate and fast taxonomic characterization of bacterial isolates from WGS data. Hence, our results suggest WGS-based genotyping to be a promising alternative to the MS-based biotyping in clinical settings. Moreover, we suggest that complementary information should be used for the evaluation of taxonomic classification tools, as public databases may suffer from suboptimal annotations.

Key words: bacteria, taxonomy, MALDI-TOF MS, whole-genome next-generation sequencing

Introduction

In the light of the global increase of antibiotic-resistant microorganisms, rapid and accurate pathogen characterization—i.e. their classification into organism groups—is essential for an effective treatment of infectious diseases [1]. This facilitates patient stratification and personalized therapies.

Several approaches have been developed for the taxonomic characterization of bacterial isolates. The classical microbiological approaches are built on a large basis of constantly revised expert knowledge and typically involve Gram staining, analysis of culture growth, phenotype and biochemical reaction patterns [2]. These methods are increasingly augmented by high-throughput molecular methods such as 16S ribosomal RNA (rRNA) gene sequencing [3]. However, the taxonomic resolution based on the 16S rRNA gene alone is limited [3, 4]. Another alternative is taxonomic analysis using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) where the obtained protein mass spectra are compared against a reference database [5]. This proteome-based approach is characterized by high accuracy [5–7], low operating costs and quick turn-around times [8, 9]. Usually, MALDI-TOF MS-based analysis is applied to pure cultures, but there are studies performing taxonomic classification of mixed microbial communities [10, 11]. The reference spectra are typically not disclosed by the manufacturers, thus hampering the study and expansion of the existing references [12]. However, several attempts to create a publicly available database exist [13, 14] with the database 'Spectra' [15], curated by the Public Health Agency of Sweden, providing a sizeable data source containing >5000 spectra of bacteria and fungi. Although various factors might negatively affect the analysis outcome, e.g. sample preparation or age, limiting the result comparability between laboratories [7, 12, 16, 17], MS-based pathogen identification is applied in many diagnostic laboratories [1].

In part, driven by decreasing costs and faster turn-around times, whole-genome sequencing (WGS) has gained importance for the identification of pathogens as well as for antimicrobial resistance analyses and outbreak monitoring [2]. The existing sequencing-based taxonomic classification tools can be organized into two groups—tools relying on specific marker genes/sequences (e.g. MetaPhlAn [18] and MetaPhyler [19]) and whole-genome-based tools. The whole-genome-based approaches assign input sequences to taxa using alignments (e.g. DIAMOND [20] and TUIT [21]), k-mer matching (e.g. CLARK [22], Kaiju [23] and Kraken [24]) or alignment-free methods (e.g. Vervier et al. [25], RAIphy [26] and PhyloPythia [27–29]). While the tools' performances are evaluated in the respective publications, these performance evaluations rely on publicly available data typically generated in independent, earlier experiments. Hence, incomplete or suboptimal annotations, e.g. because of contaminating sequences, are expected to exert negative effects on the evaluations. Importantly, the use of complementary information, such as MALDI-TOF MS-based taxonomic classification, is missing. This is of particular importance in the context of clinically relevant pathogen identification [2].

Here, we newly sequenced 846 pathogenic, gram-negative bacterial clinical isolates [including, among others, *Escherichia* spp. (22%), *Proteus* spp. (14%), *Klebsiella* spp. (16%), *Pseudomonas* spp. (11%), *Enterobacter* spp. (6%), *Salmonella* spp. (8%) and *Acinetobacter* spp. (6%)] and evaluated the classification performance of a set of WGS-based taxonomic classification tools (CLARK, DIAMOND/MEGAN, Kaiju, Kraken and TUIT). The 'ground truth' taxonomic assignments were established by confirming the expert-driven taxonomy using a Bruker Biotyper MALDI-TOF MS system. Moreover, we newly sequenced 200 *Staphylococcus aureus* isolates and performed the same analysis as for the gram-negative bacteria where the 'ground truth' comprised only the MS-based taxonomy. Our results demonstrated that certain WGS-based approaches allow for an accurate taxonomic classification, and thus, can be considered as promising alternatives to MS-based biotyping. Moreover, the complementary information of the protein mass spectra is a powerful alternative to relying on existing, yet potentially misleading, publicly available data.

Materials and methods

Bacterial isolates

Our first data set consisted of 846 gram-negative bacterial clinical isolates collected for diagnostic purposes. The isolates were characterized by microbiologists from the respective laboratory according to the institutional guidelines for routine clinical microbiological testing, which was state of the art at the time of testing (Supplementary Table S1). The overview of the taxonomic assignments was created with Krona [30] (Figure 1). The samples are part of the microbiology strain collection of Siemens Healthcare Diagnostics (West Sacramento, CA). For 240 isolates, the data set included the collection date (Supplementary Figure S1), and for 783 isolates, the collection location (country or continent) was provided (Supplementary Table S1).

The second data set included 200 *S. aureus* clinical isolates, which are part of the *S. aureus* strain collection of Saarland University Medical Center. For all isolates, the location of the isolation (country) and the isolation year (except for one sample) were provided (Supplementary Table S2, Supplementary Figure S1).

DNA extraction

Four streaks of each gram-negative bacterial isolate were cultured on trypticase soy agar containing 5% sheep blood, and cell suspensions were made in sterile 1.5 ml collection tubes containing 50 µl Nuclease-Free Water (AM9930, Life Technologies). Bacterial isolate samples were stored at −20°C until nucleic acid extraction. The Tissue Preparation System (TPS) (096D0382-R 02_01_B, Siemens) and the VERSANT® Tissue Preparation Reagents (TPR) kit (10632404B, Siemens) were used to extract DNA from these bacterial isolates. TPS for nucleic acid extraction has been described previously [31–33]. Before extraction, the bacterial isolates were thawed at room temperature and were pelleted at 2000g for 5 s. The DNA extraction protocol DNAext was used for complete total nucleic acid extraction of samples. The total nucleic acid eluates were then transferred into 96 well quantitative polymerase chain reaction (qPCR) detection plates (401341, Agilent Technologies) for RNase A digestion, DNA quantitation and plate DNA concentration standardization processes. Rnase A (AM2271, Life Technologies), which was diluted in nuclease-free water following manufacturer's instructions, was added to 50 µl of the total nucleic acid eluate for a final working concentration of 20 µg/ml. Digestion enzyme and eluate mixture were incubated at 37°C for 30 min using Siemens VERSANT® Amplification and Detection instrument. DNA from the RNase-digested eluate was quantitated using the Quant-iT™ PicoGreen dsDNA Assay (P11496, Life Technologies) following the assay kit instruction, and fluorescence was determined on the Siemens VERSANT® Amplification and Detection instrument. In total, 25 µl of the quantitated DNA eluates were transferred into a new 96 well PCR plate for plate DNA concentration standardization before library preparation. Elution buffer from the TPR kit was used to adjust DNA concentration. The standardized DNA eluate plate was then stored at −80°C until library preparation.

Pure isolates from the *S. aureus* data set were grown overnight in brain heard infusion liquid culture with regular shaking (3 ml, 150 rpm). In total, 1 ml of overnight culture was centrifuged (10 min at 5000g), and the pellet was resuspended in P1 buffer (Qiagen), supplemented with 4 µl lysostaphin (10 mg/ml, frozen stock solution, Sigma) and incubated at 37°C (30 min, 900 rpm) for enzymatic digestion of *S. aureus* cell walls. Protein K extraction was performed at 56°C (30 min) by adding 300 µl lysis buffer and protein K solution (Maxwell 16 LEV Blood DNA kit, Promega). Following automated nucleic acid isolation (Promega Maxwell) culture extracts were eluted in 75 µl nuclease-free water. Quality of high molecular DNA without DNA degradation was confirmed by standard agarose gel electrophoresis.

Next-generation sequencing

Before library preparation, quality control of isolated bacterial DNA was conducted using a Qubit 2.0 Fluorometer (Qubit dsDNA BR Assay Kit, Life Technologies) and an Agilent 2200 TapeStation (Genomic DNA ScreenTape, Agilent Technologies). Next-generation sequencing libraries were prepared in 96 well format using NexteraXT DNA Sample Preparation Kit and NexteraXT Index Kit for 96 indexes (Illumina) according to the manufacturer's protocol. The resulting sequencing libraries were quantified in a qPCR-based approach using the KAPA SYBR FAST qPCR MasterMix Kit (Peqlab) on a ViiA 7 Real-Time PCR System (Life Technologies). In total, 96 samples were pooled per lane for paired-end sequencing (2×100 bp) on Illumina Hiseq2000 or Hiseq2500 sequencers using TruSeq PE Cluster v3 and TruSeq SBS v3 sequencing chemistry (Illumina). Basic sequencing quality parameters were determined using the FastQC quality control tool for high-throughput sequence data [34], and the reports were summarized using MultiQC (version 0.8) [35] for the gram-negative isolates and *S. aureus*, respectively (Supplementary Tables S3 and S4, Supplementary Figures S2 and S3). A subset of gram-negative samples was resequenced because of low read coverage in the initial run; data of both runs were subsequently merged (Supplementary Table S1).

Proteome-based identification

Bacterial isolates were cultured on trypticase soy agar containing 5% sheep blood (BD BBL) and incubated at 35°C for 18–24 h. Isolates were subjected to MALDI-TOF MS analysis using the Bruker Biotyper 3.1.65 (Bruker Daltonics, Bremen, Germany). Isolated colonies were directly smeared onto a polished steel target plate (Bruker Daltonics). Matrix (α-cyano-4-hydroxycinnamic acid, Bruker Daltonics), reconstituted as recommended by the manufacturer (50% acetonitrile, 47.5% water and 2.5% trifluoroacetic acid), was added to the cellular material on the target plate.

Following successful calibration with the Bacterial Test Standard (Bruker Daltonics), bacterial isolates were tested on the Bruker Biotyper (flexControl version 3.3.108.0 and flexAnalysis 3.3.80.0) following the manufacturer's instructions. Mass spectra were obtained, and scores were generated. Scores of ≥2.0 were considered probable species identifications, scores of ≥1.7 but <2.0 were considered probable genus identifications and scores <1.700 were considered not reliable identifications, i.e. as not identified. The used cutoff values were defined according to the manufacturer's guidelines.

From 846 isolates analyzed with the Bruker Biotyper system, 100 samples were retested in a second run because they were not identified or yielded ambiguous classification in the first run. Best hits of the first and the second run were considered as the classification results. These included either species- or genus-level assignments with respect to the score cutoff values as described above. For all species assignments, the corresponding genus was determined using the R-package taxize [36] and the NCBI taxonomy database [37] (accessed 4 October 2016). The results of both runs were consolidated as follows: if the runs disagreed on the species level but not on the genus level, then only the genus was saved; if the runs disagreed on the genus level, the sample was considered as unclassified; and if a sample was classified at the species level in one run but only at the genus level in the other run and both genus-level assignments were concordant, then the species-level assignment was saved. Detailed information on each sample can be found in Supplementary Table S1.

The identification of the isolates from the *S. aureus* data set was performed by a MALDI-TOF MS analysis using standard protocol (Bruker Biotyper, Bruker Daltonics).

Genome-based identification of bacteria

We applied five tools for whole-genome-based taxonomic classification: BLAST-based tools DIAMOND [20] Lowest Common Ancestor (LCA) assignment using MEGAN [38]) and TUIT [21], and k-mer-based approaches CLARK [22], Kaiju [23] and Kraken [24].

CLARK: Version 1.1.3 was used, and the database was created with the respective script at the species level using finished bacterial genomes from the NCBI RefSeq database (2 November 2015). The tool was run in default mode, k-mer lengths were set to 21, 25, 29 or 31 and forward and reverse paired-end reads were used as input. Report files were created using 'getAbundance' with default parameters.

Kaiju: Version 1.4-7 was used with the default database of complete genomes downloaded from NCBI FTP server (30 June 2016). Paired-end reads were used as input with the default run mode Maximum Exact Matches (MEM). Report files on species level were created using 'kaijuReport'.

Kraken: Version 0.10.4-beta was used with the default database containing finished genomes from the NCBI RefSeq database (13 January 2015). The k-mer lengths were set to 21, 25, 29 or 31, and forward and reverse paired-end reads were used as input. Report files were created from the raw output using 'kraken-report'.

DIAMOND/MEGAN: As DIAMOND has no direct support for paired-end reads, we used only forward reads as input. Version 0.6.13.48 was used with BLASTX search against the NCBI nonredundant protein sequence database (nr) (27 February 2015) with default parameters. The output was further processed using the LCA method implemented in MEGAN (version 5.10.6, tool blast2lca with default parameters and GenInfo Identifier (GI) number taxon mapping from March 2015) and summarized by counting the number of mapped sequences for each listed taxon.

TUIT: As in the case of DIAMOND, only forward reads were used as input. Furthermore, the FASTQ file was converted to FASTA format using FASTX-Toolkit [39] (version 0.0.14, with '-Q 33'), and unique reads were collapsed to reduce the number of input sequences. TUIT (version 1.0.3.2) was used with local BLAST search against the NCBI nucleotide collection (nt) (4 April 2014) and default parameters. The output was summarized as described for DIAMOND/MEGAN.

Result summaries

From the individual tools' reports, the following information was computed: species taxon with maximal percentage of mapped sequences (first best species hit with respect to all reported sequences), the number of sequences mapped to the best hit species taxon, the number of sequences classified at the species level and the number of sequences mapped to the expected species taxon, i.e. taxa obtained by the merged MS-based analysis results in case of the gram-negative isolates. The total number of sequences was set to the number of input sequences, i.e. the number of reads for the CLARK, DIAMOND/MEGAN, Kaiju and Kraken results, and the number of FASTA sequences after converting FASTQ files into FASTA files for the results of TUIT. The summarized results can be found in Supplementary Tables S1 and S2.

Performance measures

For each WGS-based summary file, the following performance measures were calculated: the sensitivity, precision and F-measure values with respect to the best species hit and expected species taxon (Supplementary Table S1). Sensitivity was defined as the ratio of reads assigned to the species taxon and the total number of reads. Precision was defined as the ratio of reads assigned to the species taxon and reads classified at the species level (i.e. assigned to any species taxon). F-measure was defined as 2 × (sensitivity × precision)⁄(sensitivity + precision).

Runtime analysis

To compare the runtimes of the tools, we randomly selected five gram-negative samples whose number of reads was between 100 000 and 1 000 000 to reduce computational cost. For each sample and each tool, the elapsed (wall clock) time was estimated three times using 'GNU time' (version 1.7). Before measuring the time, the tools were 'pre-run' on a single sample. The tools were called in the same way as described above with the following additional settings: the number of threads was set to 30 using the parameter '–threads' for Kraken and DIAMOND, '-n' for CLARK and '-z' for Kaiju. For TUIT, only the number of threads for the BLAST search can be set manually through the parameter 'NumThreads' in the supplied property file. This parameter was also set to 30. Furthermore, we used the option '–preload' for Kraken. The analysis was performed on a server with 500 Gb RAM and 64 processors [AMD Opteron™ Processor 6378] with 1400 MHz. The time spent on downloading and creating the databases was not considered. The final runtime per sample was computed as the mean over the three repetitions and normalized by the number of read pairs in the corresponding FASTQ files.

Effect of read processing on results of CLARK and Kraken

CLARK and Kraken were additionally run on paired reads pre-processed with Trimmomatic [40] as described below. K-mer length was set to 31. All other parameters and output processing were kept as described above.

Identification of species contained in the reference databases

Whether the expected species were represented by the used reference databases of the WGS-based tools was determined as follows: for CLARK, Kaiju, Kraken and TUIT, the GI numbers were extracted from used nucleotide sequences and mapped to the taxonomy names using the taxonomy mapping files of NCBI; then we checked whether the expected taxa were contained in the set of retrieved taxonomy names; for DIAMOND/MEGAN, the sequence titles were extracted from the used nr database, filtered to retrieve only those with one unique taxonomy name and used to search for the expected taxa.

Identification of candidate mixtures

To detect samples containing genomic data of more than one organism, we performed a homogeneity analysis based on the WGS data. The raw reads were trimmed using Trimmomatic [40] (version 0.35, command line parameters: PE ILLUMINACLIP:NexteraPE-PE.fa:1:50:30 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36), assembled into scaffolds using SPAdes [41] (version 3.6.2, command line parameters: -k 21,33,55 –careful) and annotated by PROKKA [42] (version 1.11, command line parameters: –gram neg –mincontiglength 200). The homogeneity of individual samples was assessed using a set of 'essential genes'. These genes are in single copy and conserved in 95% of all sequenced bacteria [43]. A sample was considered a candidate mixture if >10 of the essential genes were found in multiple copies (Supplementary Tables S1 and S2).

Multilocus sequence typing

Multilocus sequence typing (MLST) profiles for each isolate were obtained using a pipeline implemented by Torsten Seeman [44] (version 2.6) and PubMLST schemes [45] (8 August 2016). The scaffold FASTA files obtained by PROKKA (mentioned earlier) were used as input. The output contained the closest PubMLST scheme, the corresponding sequence type (ST; if available) and the allele IDs (Supplementary Table S3). Results containing more than one missing allele were considered as unreliable.

Validation set

A validation set of gram-negative isolates used to access the performance of WGS-based results was defined as follows. It included only samples identified at species level by MS-based analysis and whose species taxon was supported by the expert-driven taxonomy. If the expert-driven taxonomy contained two species taxa, then it was sufficient that one of them was the same as the MS-based taxon. If the expert-driven taxonomy included only genus, then it was required to be the same as the genus MS-based species taxon. Furthermore, samples were excluded if their assembly failed or if they had <200 000 reads or if they were identified as candidate mixtures (Supplementary Table S1) resulting in 656 samples in total. For *S. aureus*, the validation set contained 194 isolates after removing samples because of failed assembly or if they had <200 000 reads or if they were identified as candidate mixtures (Supplementary Table S2).

Results

Gram-negative isolates

Our data set consisted of 846 gram-negative bacterial clinical isolates collected and identified by microbiologists for diagnostic purposes (Figure 1). These isolates were part of the microbiology strain collection at Siemens Healthcare Diagnostics (West Sacramento, CA). We sequenced whole genomes of all 846 isolates using an Illumina HiSeq system and performed WGS-based taxonomic classification using in silico methods. Additionally, a Bruker Biotyper MALDI-TOF MS system was used for taxonomic characterization where a subset of 100 isolates was reclassified in a second run. For 240 isolates, the date of collection was available covering a time span from 1986 to 2011 (Supplementary Figure S1). Furthermore, for 783 isolates, the data set included the collection location: a majority of samples was collected in North America (738), followed by Japan (23), Europe (21) and Australia (1).

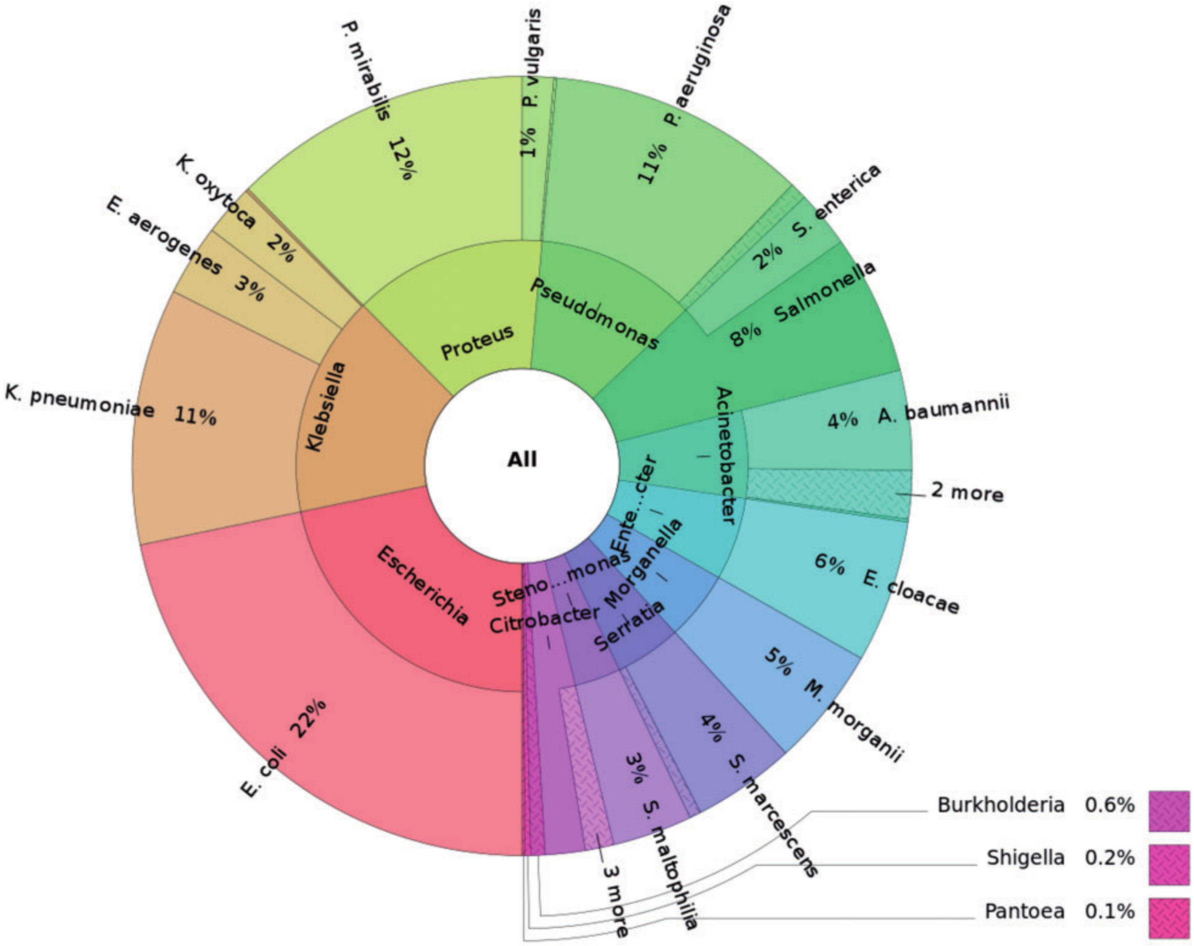


Figure 1. Taxonomic composition of the 846 gram-negative isolates based on expert-driven taxonomy.

In the MS-based analysis of the 846 isolates, eight (0.9%) samples in the first run and four from the 100 reanalyzed samples (4%) were not identified. Only two samples remained unclassified after both runs. In total, 734 of 846 (86.8%) and 86 of 100 (86%) isolates were resolved to the species level in the first and second run, respectively, while 104 of 846 (12.3%) and 10 of 100 (10%) were classified at the genus level only. The score values varied from 1.73 to 2.57 (Supplementary Figure S4). For the 74 samples identified at the species level in both MS runs, 52 (70.3%) had concordant results. From the remaining 22 samples with divergent species taxa, 16 were assigned to the same genus. Among these, 11 samples were identified as *Serratia ureilytica* in the first run but reclassified as *Serratia marcescens* in the second run. A stronger concordance of 83 of 90 samples (92.2%) was observed at the genus level. After merging the results of both runs, 723 from 846 samples (85.5%) were classified at the species level, 114 (13.5%) at the genus level only and 9 (1.1%) were considered as unclassified (Supplementary Table S1).

Next, we examined the agreement between the MS-based results and the expert-driven taxonomy over all isolates (846). Concordant species assignments (in case of two taxa, one match was sufficient) were observed in 646 (76.4%) cases

(Figure 2, Supplementary Figure S5). Moreover, the MS-based results were supported by all WGS-based tools (for CLARK and Kraken, *k*=31) in 602 (71.2%) cases where 24 (2.8%) assignments were not supported by the expert-driven taxonomy. The validation set for the evaluation of the WGS-based tools was defined comprising 656 isolates with verified species classification (Supplementary Table S1). The taxonomy of the selected isolates was additionally confirmed by the assembly-based MLST results. For species with available MLST schemes in the PubMLST database, for all samples, except one *Klebsiella pneumoniae* isolate, the closet reported scheme belonged to the expected species taxon (Supplementary Figure S6), and in 354 of 458 cases, a ST could be assigned (Supplementary Table S1). In the following, if not stated otherwise, only the isolates from the validation set were used for the analysis of the WGS-based results.

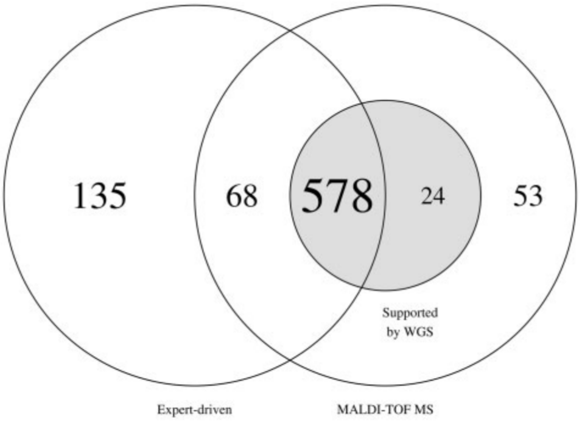
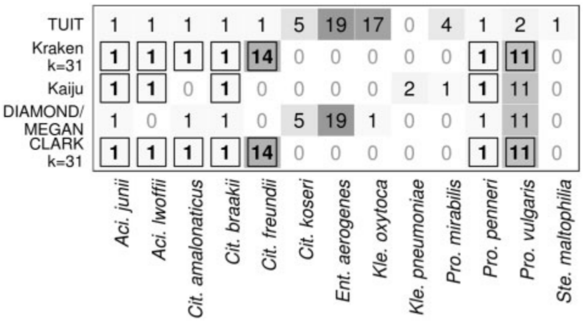


Figure 2. Euler diagram of the species taxa of the expert-driven and MS-based identifications. In cases where two species taxa were given by the expert-driven taxonomy, a match to one of the taxa was sufficient to report concordance. The third set lying within the MS-based taxa set represents assignments supported by CLARK (31-mers), Kaiju, DIAMOND/MEGAN, Kraken (31-mers) and TUIT. Its difference with the set of the expert-driven species taxa includes three isolates without an expert-based species assignment.



	Aci. junii	Aci. lwoffii	Cit. amalonaticus	Cit. braakii	Cit. freundii	Cit. koseri	Ent. aerogenes	Kle. oxytoca	Kle. pneumoniae	Pro. mirabilis	Pro. penneri	Pro. vulgaris	Ste. maltophilia
TUIT	1	1	1	1	1	5	19	17	0	4	1	2	1
Kraken k=31	**1**	**1**	**1**	**1**	**14**	0	0	0	0	0	**1**	**11**	0
Kaiju	**1**	**1**	0	**1**	0	0	0	0	2	1	**1**	11	0
DIAMOND/ MEGAN	1	0	1	1	0	5	19	1	0	0	1	11	0
CLARK k=31	**1**	**1**	**1**	**1**	**14**	0	0	0	0	0	**1**	**11**	0

Figure 3. A number of misclassified samples from the validation set per species for CLARK (31-mers), Kaiju, DIAMOND/MEGAN, Kraken (31-mers) and TUIT. Only the expected species taxa for which at least one tool yielded a misclassification were included. The number of misclassifications is provided within each cell—the higher the value, the darker the background color. For CLARK, Kaiju and Kraken, the numbers of species missing in the used databases are printed in bold and are highlighted by a black rectangle. The genus taxa were abbreviated by the first three letters.

CLARK and Kraken require the k-mer length to be fixed when building the respective reference databases. In both cases, the values was set to 31, the default value of Kraken. In an additional analysis, we confirmed the observation made for Kraken and CLARK that higher k-mer values are associated with higher precision and lower values with higher sensitivity [22, 24] (Supplementary Figure S7). We examined the results obtained from CLARK, DIAMOND/MEGAN, Kaiju, Kraken and TUIT regarding the wrongly assigned species taxa and their presence in the reference data sets of the used tools (Figure 3). The numbers of incorrect classifications per species were comparable among all five tools with some exceptions: 5 *Citrobacter koseri* and 19 *Enterobacter aerogenes* samples were misclassified by DIAMOND/MEGAN and TUIT, whereas

CLARK, Kaiju and Kraken yielded no misclassifications for these species; TUIT assigned all 17 *Klebsiella oxytoca* samples to *K. pneumoniae* but had only two wrong assignments for *Proteus vulgaris*, whereas the other tools misclassified 11 *P. vulgaris* samples. For CLARK and Kraken, in all 30 cases of incorrect species classifications, the expected taxon was not contained in the respective reference databases. For Kaiju, this was the case for only 4 of the 18 incorrect assignments. The databases used for DIAMOND and TUIT contained all expected species taxa. To perform a fair comparison of the tools, only isolates belonging to species included in the reference data sets of all five tools (626 samples) were considered if not stated otherwise.

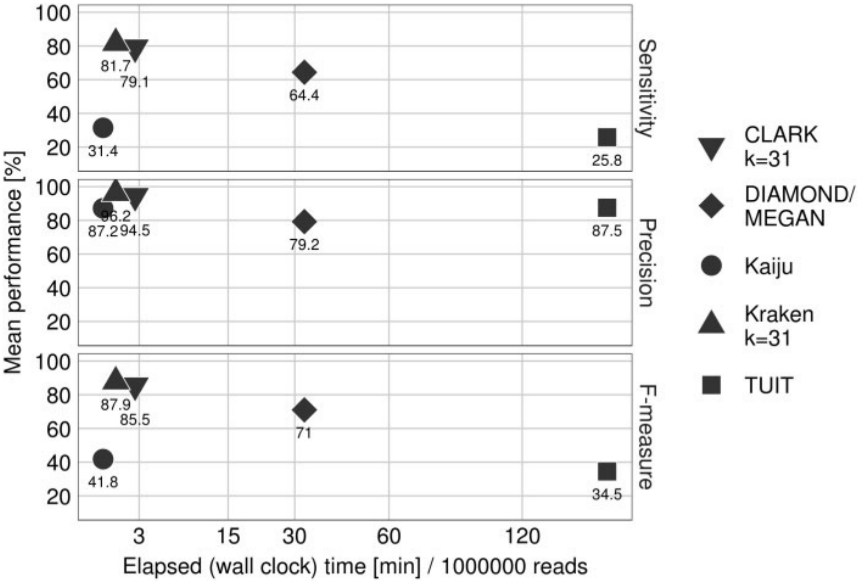


Figure 4. The mean runtime (*n*=5) per 1 million reads measured using five randomly chosen gram-negative samples, and the mean sensitivity, precision and F-measure percentages computed with respect to the expected species taxa for CLARK (31-mers), DIAMOND/MEGAN, Kaiju, Kraken (31-mers) and TUIT. Only samples from the validation set and with expected species present in all used reference databases were used. The x-axis is square root transformed.

The overlap of species assignments between all four tools was 92.3% (578 of 626 samples) (Supplementary Figure S8). CLARK and Kraken obtained same species assignments for all samples, while the lowest concordance was found between Kaiju and TUIT (92.3%). As next, we compared the results of CLARK, DIAMOND/MEGAN, Kraken and TUIT with respect to the expected species taxa. For each of the five tools, the percentage of correct species- and genus-level assignments was computed (Supplementary Figure S9). The best results were obtained by CLARK and Kraken with 100% of correctly assigned species taxa followed by Kaiju (99.5%), DIAMOND/MEGAN (96%) and TUIT (92.7%). We then considered the mean classification performance computed with respect to the expected species taxa. CLARK and Kraken demonstrated comparable mean sensitivity values of 79.1 and 81.7%, respectively; DIAMOND/MEGAN had a mean sensitivity of 64.6 % followed by Kaiju (31.4%), and TUIT (25.8%) (Figure 4). The highest mean precision was achieved by Kraken (96.2%) followed by CLARK (94.5%), TUIT (87.5%), Kaiju (87.2) and DIAMOND/MEGAN (79.2%). Regarding the F-measure, best mean performance was achieved by Kraken (87.9%) followed by CLARK (85.5%), DIAMOND/MEGAN (71%), Kaiju (41.8) and TUIT (34.5%). Finally, we examined the distribution of the best hit performance values for correctly and wrongly classified isolates (Supplementary Figure S10). We observed that isolates assigned to wrong species taxa tended to have a combination of lower sensitivity and precision values compared with correctly classified isolates. However, there was no clear separation of both groups. Moreover, it should be noted that the pairwise sensitivity and precision values of DIAMOND's results were often almost identical.

Besides the classification performance, computational runtime is a relevant factor in choosing which software to use, especially when analyzing large-scale data sets. Hence, we compared the runtimes of the tools based on five randomly selected samples (Figure 4). Among the herein used tools, TUIT was the slowest with an average of 169 min per 1 million reads followed by DIAMOND (32.6 min). CLARK, Kaiju and Kraken achieved the best results requiring <3 min per 1 million reads, with Kaiju being the fastest (<1 min).

Finally, we investigated whether adapter and quality trimming of the raw reads would adversely affect the classification results with the focus on CLARK and Kraken, as they demonstrated comparable results and achieved better performance than the other tools. We compared the results obtained using raw or processed reads (for *k*=31). Regarding the best species hits of all samples from the validation set (656), all assignments stayed consistent when using CLARK and Kraken. Moreover, we compared the percentages of raw and trimmed reads assigned to the best hit and expected species taxa, respectively. The distribution of the absolute differences (Supplementary Figure S11) was inspected, and the 99th percentile (at the value of 2.6 for CLARK and 2.5 for Kraken) was defined as the cutoff for outlier detection: The numbers of outliers for the best hits and expected species taxa, respectively, were seven for CLARK and Kraken.

Staphylococcus aureus isolates

We whole genome sequenced 200 *S. aureus* clinical isolates from the *S. aureus* strain collection of Saarland University Medical Center. The samples were collected between 1976 and 2014 (Supplementary Figure S1). The majority of the isolates (187) was collected from Germany, 11 were from Mozambique, 1 from Switzerland and 1 from the United States (Supplementary Table S2). In the assembly-based MLST analysis, for all isolates, except one, in the validation set, the closest reported MLST scheme belonged to *S. aureus*, and 173 isolates were assigned to known MLST profiles.

We performed an analogous analysis as for the gram-negative isolates to compare the WGS-based results. The tools were concordant with the assignments of all isolates in the validation set except for one case where Kaiju disagreed with the other tools (Supplementary Figure S12). Only one isolate was not assigned to *S. aureus* by CLARK, DIAMOND/MEGAN, Kraken and TUIT; two isolates were misclassified by Kaiju (Supplementary Figure S13). CLARK and Kraken achieved high mean sensitivity values of 90.8% and 91.3%, respectively (Supplementary Figure S14). Lower values were observed for TUIT (77.1%), DIAMOND/MEGAN (75.9%) and Kaiju (10.9%). CLARK, DIAMOND/MEGAN, Kraken and TUIT demonstrated high-precision values of >97%, whereas Kaiju had a lower value of 85.9%. Accordingly, CLARK and Kraken had highest F-measure values of 94.4 and 94.7%, respectively, followed by TUIT (86.5%), DIAMOND/MEGAN (85.3%) and Kaiju (18.9%).

Discussion

Rapid and accurate pathogen characterization is essential for an effective treatment of infections facilitating patient stratification and personalized therapies. WGS is gaining importance in the analysis of bacterial cultures derived from patients with infectious diseases. Various computational approaches have been developed to perform taxonomic analysis based on WGS data. However, evaluations using newly sequenced clinical samples and complementary information confirming the taxonomy are missing. Here, we performed WGS-based taxonomic analysis of 846 gram-negative bacterial isolates and validated the results by comparing with MS-based classifications obtained using a Bruker Biotyper MALDI-TOF MS system and confirmed by expert-driven taxonomy. Our data set included species which are frequently found to be responsible for nosocomial (hospital-acquired) infections, such as *Acinetobacter baumannii*, *Escherichia coli*, *Klebsiella* spp. and *Pseudomonas aeruginosa* [46]. Additionally, we included a data set of 200 *S. aureus* isolates.

To determine the concordance between the expert-driven and MS-based taxonomy of the gram-negative isolates, we first analyzed the MS-based results to determine samples with uncertain or ambiguous classifications. In general, possible limitations of MS-based analysis include, but are not restricted to, the limited differentiation of *E. coli* and some *Shigella* spp. Isolates [7, 9, 47–49], and also inaccurate differentiation of species in other groups such as *Acinetobacter* [48], *Citrobacter* [50] and *Enterobacter cloacae* complex [48] and the missing identification of *Salmonella* isolates below the genus level [51]. In our MS-based analysis, most samples (about 86%) were classified at the species level and the taxa of samples classified only at the genus level included, among others, the genera imposing particular challenges to MS-based biotyping as mentioned above (Supplementary Table S1). We found that both MS runs produced concordant results in 52 of 74 (70.3%) and 84 of 90 (93.3%) cases at the species and at the genus level, respectively. From the 22 discordant species-level assignments, 11 isolates were first identified as *S. ureilytica* and then reclassified as *S. marcescens*. *S. ureilytica* is a relatively new species [52] whose identification by a MALDI-TOF MS system was shown to be challenging [50, 53], potentially explaining the observed ambiguities. In the final taxonomic assignment, the reported inconsistencies were resolved such that 723 samples were classified at the species level, 115 at the genus level only and 8 samples were considered as unclassified because of divergent assignments or failed identification. We then compared the resulting species taxa with the expert-driven results and observed a high concordance of 76.4%. However, in ca. 3% of the cases, all tested WGS-based tools were concordant with the MS-based result, which was not supported by the expert-driven classification. We could assume that in these cases, the expert-driven taxonomy was incomplete (no species taxon) or incorrect, which demonstrates the importance of using complementary information to confirm the identification results.

Subsequently, we defined a validation set including only samples with confirmed taxonomy to be used for the evaluation of the WGS-based results. For any taxonomic classification approach, the availability and quality of the reference data are crucial for an accurate identification. Sequence-based tools for taxonomic classification generally rely on publicly available data sources. CLARK and Kraken construct their k-mer databases from finished genomes of the NCBI RefSeq database [54, 55], which included 2785 bacterial data sets (containing chromosome and/or plasmid DNA sequences) in this study. DIAMOND and TUIT use the NCBI nonredundant protein (nr) and nucleotide (nt) collections [56], respectively. The NCBI nr collection includes data from GenBank, RefSeq, UniProtKB/Swiss-Prot, PDB and PRF; and the nt collection includes data from RefSeq and GenBank except EST, GSS, STS and HTG [56]. Kaiju can use complete genomes from the NCBI RefSeq database or the nr collection. In this study, Kaiju was run using proteins extracted from 5135 complete genomes. Considering the presence of expected species in the different reference data sets, in only few cases, the availability of the respective species was sufficient for correct classifications (Figure 3). Only TUIT was able to correctly identify most of the *P. vulgaris* isolates, though this species was also contained in the reference data sets of DIAMOND and Kaiju. For the single *P. vulgaris* genome used by Kaiju (NZ_CP012675.1), we found that the nucleotide sequence of the rpoB gene (DNA-directed RNA polymerase sub-unit beta, WP_004246906.1), shown to be appropriate to distinguish *Proteus* spp. [57], was 100% identical to the rpoB gene from the complete *Proteus mirabilis* genome (NZ_CP012674.1, ALE23450). Furthermore, the average nucleotide identity value [58] (http://enve-omics.ce.gatech.edu/ani/index) of both genomes was >99.9%. Based on these findings, we assumed that NZ_CP012675.1 was misclassified explaining why Kaiju was not able to correctly identify the *P. vulgaris* isolates. We hypothesized that similar reasons may be responsible for other incorrect assignments where the expected species was present in the used database but whose isolates were nevertheless misclassified. Focusing only on species presumably contained in all databases, our analysis demonstrated that WGS-based identification approaches can yield highly accurate results. All tools classified >92% (best result 100%) of the gram-negative samples correctly. CLARK and Kraken demonstrated best mean sensitivity about 80% followed by DIAMOND/MEGAN with 64.4%. Kaiju and TUIT had a comparably low mean sensitivity (31.4 and 25.8%) but better precision (87.2 and 87.5%) than DIAMOND/MEGAN (79.2%). The highest precision was observed for Kraken (96.2%) and CLARK (94.5%). The low sensitivity of TUIT may be because of missing matches to the reference database or because the default TUIT cutoff values used during the assignment of the LCAs, were too strict. The authors of TUIT suggest a trial-and-error procedure to adjust the cutoffs [21], but this was prohibited by the high computational runtime of the tool. In contrast, the default parameters used for assignment filtering and LCA assignment with MEGAN for the DIAMOND results appeared to be too permissive, as the sensitivity and precision values were close to each other. Furthermore, it should be noted that CLARK, Kaiju and Kraken may have benefited from using paired-end data, while DIAMOND and TUIT were run on forward reads only. Moreover, DIAMOND and Kaiju are only able to classify protein-coding sequences, as the reads are matched to protein databases affecting their sensitivity. Though we observed a tendency of lower performance values for incorrectly classified isolates, some wrongly and correctly assigned isolates demonstrated comparable sensitivity and precision values. We hypothesized that these cases may either include not detected contaminated isolates or isolates belonging to a species missing in the database but closely related to other species of the same genus with available reference data. These isolates would require a closer examination, e.g. considering all taxa exceeding a reasonable abundance cutoff. The overall results demonstrate the importance of a comprehensive and representative reference database for a successful and precise taxonomic classification, which is even more crucial within a clinical setting.

The performance of the WGS-based tools on the *S. aureus* data set was comparable with the observations made for the gram-negative isolates: at least 99% of all samples were classified as *S. aureus*; the exceptions were one sample (ID 191) classified as *Enterococcus faecium* by Kaiju and one sample (ID 80) identified as *Staphylococcus carnosus* by all five tools. In the latter case, we could assume that the MS-based taxon was wrong or that a wrong probe was used for WGS. Highest sensitivity was achieved by CLARK and Kraken (>90%), and all tools except Kaiju had a mean precision >97%.

Considering the runtime of the tested WGS-based tools, CLARK, Kaiju and Kraken were substantially faster than DIAMOND and TUIT requiring only a few minutes to process a million of reads. Furthermore, we also evaluated the robustness of CLARK and Kraken with respect to adapter and quality trimming of the raw reads and observed that this procedure had only marginal effects on the classification results.

In summary, the k-mer and exact matching-based tools CLARK and Kraken appear to be the primary choices with respect to classification performance, usability and runtime among the herein tested approaches. Kaiju represents an appealing alternative, as it was faster than CLARK and Kraken, and requires no parameters to be set to create a reference database. But it should be kept in mind that the tool can classify only protein-coding sequences. Overall, taxonomic classification of bacterial isolates based on WGS data provides highly accurate results, and thus, represents a promising alternative to MS-based biotyping. WGS would also enable further analyses, such as phylogenetic analysis and genotyping, which are mandatory for surveillance of outbreaks and antimicrobial resistance. However, multiple issues have to be addressed before WGS-based approaches can be applied in clinical settings. The effect of different library preparation and sequencing methods on the quality of the identification results should be investigated and quantified. As the comprehensiveness and the quality of the reference database has a high impact on the reliability of the taxonomic assignments, a careful selection and validation of the reference data would be necessary. This holds especially for organisms represented by only one genome as seen in case of the (most likely mislabeled) *P. mirabilis* genome used by Kaiju. Large-scale efforts, such as the ‘Genomic Encyclopedia of Bacteria and Archaea’ [59] and its pilot studies, and the ‘100K Foodborne Pathogen Genome Project’ [60] are expected to greatly expand the volume and diversity of available reference genomes. An additional aspect is the genomic variability of bacteria and in particular the differences between pathogenic and nonpathogenic species. The lifestyle of a bacterium influences to a great extent its genome size and variability. Pathogenic species represent specialized organisms leading an allopatric lifestyle and are characterized by a significantly reduced genome compared with species from a sympatric environment [61, 62]. Furthermore, the genome of a bacterial strain can be seen as a composition of ‘core’ genes, conserved among many strains of the same species, and ‘accessory’ genes, which vary between different strains [62]. The set of all genes found in a species is referred to as a pan-genome [63]. The core genome similarity is considered to be a good approach to define bacterial species relevant for humans [62]. However, it has also been proposed to apply pan-genome analysis to redefine bacterial species [64, 65].

Another important point is the fact that many bacterial organisms cannot be grown in the laboratory, thus challenging their identification [66]. Single-cell sequencing is considered to be a promising solution for this problem [67]. In our analysis, we focused on cultured isolates as their accurate identification can be seen as a necessary requirement for future, culture-independent studies. Regarding the underlying concept of the classification tools, we focused in this study on alignment-based methods. But there also exist alignment-free approaches (e.g. PhyloPythia/S/S+ [27–29], RAIphy [26] and the approach of Vervier et al. [25]), and approaches combining alignment-based and alignment-free similarity measures (e.g. Borozan et al. [68]). Finally, exhaustive testing procedures using high-quality validation data should be performed including relevant human pathogens to access the reliability and accuracy of the implemented method.

Key Points

- Kmer-based taxonomic information derived from WGS data allows for accurate and fast classification of bacterial clinical isolates at species level, and is thus, an appealing alternative to MS-based analysis.
- Establishing a high-quality reference database as well as its continuous extension is vital for the correctness of the taxonomic classifications.
- The evaluation of taxonomic classification tools should include complementary information to confirm the taxonomy of the underlying data.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgement

The authors would like to thank Siemens Healthcare and Curetis GmbH for the support as well as the provided data set. The authors also thank Mathias Herrmann for comments that greatly improved the manuscript.

Funding

Siemens Healthcare and in parts by the Best Ageing (grant number 306031) from the European Union.

Availability of WGS data

The raw WGS data are available on a reasonable request for noncommercial use after signing a nondisclosure agreement.

References

1. Greatorex J, Ellington MJ, Köser CU, et al. New methods for identifying infectious diseases. Br Med Bull 2014;112:27–35. doi: 10.1093/bmb/ldu027

2. Didelot X, Bowden R, Wilson DJ, et al. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 2012;13:601–12. doi: 10.1038/nrg3226

3. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 2007;45:2761–4. doi: 10.1128/JCM. 01228-07

4. Rajendhran J, Gunasekaran P. Microbial phylogeny and diver-

sity: small subunit ribosomal RNA sequence analysis and be-

yond. Microbiol Res 2011;166:99–110. doi: 10.1016/j.micres.2010.02.003

5. van Veen SQ, Claas ECJ, Kuijper EJ. High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization-time of flight mass spectrometry in conventional medical microbiology laboratories. J Clin Microbiol 2010;48:900–7. doi: 10.1128/JCM.02071-09

6. Seng P, Drancourt M, Gouriet F, et al. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis 2009;49:543–51. doi: 10.1086/600885

7. Bizzini A, Durussel C, Bille J, et al. Performance of matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of bacterial strains routinely isolated in a clinical microbiology laboratory. J Clin Microbiol 2010;48:1549–54. doi: 10.1128/JCM.01794-09

8. Köser CU, Ellington MJ, Cartwright EJP, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 2012;8:e1002824. doi: 10.1371/journal.ppat.1002824

9. Croxatto A, Prod'hom G, Greub G, et al. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS Microbiol Rev 2012;36:380–407. doi: 10.1111/j.1574-6976.2011.00298.x

10. Mahé P, Arsac M, Chatellier S, et al. Automatic identification of mixed bacterial species fingerprints in a MALDI-TOF mass-spectrum. Bioinformatics 2014;30:1280–6. doi: 10.1093/bioinformatics/btu022

11. Zhang L, Smart S, Sandrin TR. Biomarker- and similarity coefficient-based approaches to bacterial mixture characterization using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS). Sci Rep 2015;5:15834. doi: 10.1038/srep15834

12. Sandrin TR, Demirev PA. Using mass spectrometry to identify and characterize bacteria. Microb 2014;9:23–9.

13. Mazzeo MF, Sorrentino A, Gaita M, et al. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for the discrimination of food-borne microorganisms. Appl Environ Microbiol 2006;72:1180–9. doi: 10.1128/AEM.72.2.1180-1189.2006

14. Böhme K, Fernández-No IC, Barros-Velázquez J, et al. SpectraBank: an open access tool for rapid microbial identification by MALDI-TOF MS fingerprinting. Electrophoresis 2012;33:2138–42. doi: 10.1002/elps.201200074

15. Spectra—Extended spectra database for microorganism identification by MALDI-TOF MS. http://spectra.folkhalsomyndigheten.se/spectra/. (5 April 2016, date last accessed)

16. Nicolau A, Sequeira L, Santos C, Mota M. Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF MS) applied to diatom identification: influence of culturing age. Aquat Biol 2014;20:139–44. doi: 10.3354/ab00548

17. Veloo ACM, Elgersma PE, Friedrich AW, et al. The influence of incubation time, sample preparation and exposure to oxygen on the quality of the MALDI-TOF MS spectrum of anaerobic bacteria. Clin Microbiol Infect 2014;20:O1091–7. doi: 10.1111/1469-0691.12644

18. Segata N, Waldron L, Ballarini A, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012;9:811–4. doi: 10.1038/nmeth.2066

19. Liu B, Gibbons T, Ghodsi M, et al. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics 2011;12 (Suppl 2):S4. doi: 10.1186/1471-2164-12-S2-S4

20. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2014;12:59–60. doi: 10.1038/nmeth.3176

21. Tuzhikov A, Panchin A, Shestopalov VI. TUIT, a BLAST-based tool for taxonomic classification of nucleotide sequences. Biotechniques 2014;56:78–84. doi: 10.2144/000114135

22. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 2015;16:236. doi: 10.1186/s12864-015-1419-2

23. Menzel P, Lee Ng K, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 2016;7:11257.

24. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;15:R46. doi: 10.1186/gb-2014-15-3-r46

25. Vervier K, Mahé P, Tournoud M, et al. Large-scale machine learning for metagenomics sequence classification. Bioinformatics 2016;32:1023–32. doi: 10.1093/bioinformatics/btv683

26. Nalbantoglu OU, Way SF, Hinrichs SH, et al. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics 2011;12:41. doi: 10.1186/1471-2105-12-41

27. McHardy AC, Martín HG, Tsirigos A, et al. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 2007;4:63–72. doi: 10.1038/nmeth976

28. Patil KR, Roune L, McHardy AC, et al. The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. PLoS One 2012;7:e38581. doi: 10.1371/journal.pone.0038581

29. Gregor I, Dröge J, Schirmer M, et al. PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. Peer J 2016;4:e1603.

30. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a web browser. BMC Bioinformatics 2011;12:385. doi: 10.1186/1471-2105-12-385

31. Vajapey U, Ying A, Hennig G, Li H. Validation of a fully-automated nucleic acid extraction method for fresh frozen tissue using the Siemens tissue preparation solution. Assoc Mol Pathol 2013;15:843–945.

32. Guettouche T, Rantus J, Slosek K, et al. A workflow enabling whole exome and whole genome sequencing of formalin fixed paraffin embedded samples with minimal amounts of DNA. Adv Genome Biol Technol 2013. https://www.kapabiosystems.com/assets/Guettouche_A-Workflow-Enabling-Whole-Exome-and-Whole-Genome-Sequencing-of-FFPE-Samples-with-Minimal-Amounts-of-DNA_AGBT_2013.pdf (15 November 2016, date last accessed).

33. van Eijk R, Stevens L, Morreau H, van Wezel T. Assessment of a fully automated high-throughput DNA extraction method from formalin-fixed, paraffin-embedded tissue for KRAS, and BRAF somatic mutation analysis. Exp Mol Pathol 2013;94:121–5. doi: 10.1016/j.yexmp.2012.06.004

34. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (31 May 2016, date last accessed).

35. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;btw354. doi: 10.1093/bioinformatics/btw354

36. Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R. F1000Research 2013;2:191. doi: 10.12688/f1000research.2-191.v2

37. Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2009;37:D5–15. doi: 10.1093/nar/gkn741

38. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res 2007;17:377–86. doi: 10.1101/gr.5969107

39. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/. (27 October 2015, date last accessed).

40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20. doi: 10.1093/bioinformatics/btu170

41. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77. doi: 10.1089/cmb.2012.0021

42. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9. doi: 10.1093/bioinformatics/btu153

43. Dupont CL, Rusch DB, Yooseph S, et al. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. isme J 2012;6:1186–99. doi: 10.1038/ismej.2011.189

44. Seemann T. MLST pipeline. https://github.com/tseemann/mlst. (8 August 2016, date last accessed).

45. PubMLST. www.pubmlst.org. (8 August 2016, date last accessed).

46. Peleg AY, Hooper DC. Hospital-acquired infections due to gram-negative bacteria. N Engl J Med 2010;362:1804–13. doi: 10.1056/NEJMra0904124

47. Murray PR. What is new in clinical microbiology-microbial identification by MALDI-TOF mass spectrometry: a paper from the 2011 William Beaumont Hospital Symposium on molecular pathology. J Mol Diagn 2012;14:419–23. doi: 10.1016/j.jmoldx.2012.03.007

48. Patel R. MALDI-TOF MS for the diagnosis of infectious diseases. Clin Chem 2015;61:100–11. doi: 10.1373/clinchem.2014.221770

49. Du Z, Li L, Chen C-F, et al. G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. Nucleic Acids Res 2009;37:W345–9.

50. Eigner U, Holfelder M, Oberdorfer K, et al. Performance of a matrix-assisted laser desorption ionization-time-of-flight mass spectrometry system for the identification of bacterial isolates in the clinical routine laboratory. Clin Lab 2009;55:289–96.

51. Chen JH, Ho P-L, Kwan GS, et al. Direct bacterial identification in positive blood cultures by use of two commercial matrix-assisted laser desorption ionization-time of flight mass spectrometry systems. J Clin Microbiol 2013;51:1733–9. doi: 10.1128/JCM.03259-12

52. Bhadra B, Roy P, Chakraborty R. Serratia ureilytica sp. nov., a novel urea-utilizing species. Int J Syst Evol Microbiol 2005;55:2155–8. doi: 10.1099/ijs.0.63674-0

53. Seng P, Abat C, Rolain JM, et al. Identification of rare pathogenic bacteria in a clinical microbiology laboratory: impact of matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol 2013;51:2182–94. doi: 10.1128/JCM.00492-13

54. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005;33:D501–4. doi: 10.1093/nar/gki025

55. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2011;40:D130–5. doi: 10.1093/nar/gkr1079

56. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2013;41:D8–20. doi: 10.1093/nar/gks1189

57. Giammanco GM, Grimont PA, Grimont F, et al. Phylogenetic analysis of the genera Proteus, Morganella and Providencia by comparison of rpoB gene sequences of type and clinical strains suggests the reclassification of Proteus myxofaciens in a new genus, Cosenzaea gen. nov., as Cosenzaea myxofaciens comb. nov. Int J Syst Evol Microbiol 2011;61:1638–44. doi: 10.1099/ijs.0.021964-0

58. Goris J, Konstantinidis KT, Klappenbach JA, et al. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 2007;57:81–91. doi: 10.1099/ijs.0.64483-0

59. Wu D, Hugenholtz P, Mavromatis K, et al. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. Nature 2009;462:1056–60. doi: 10.1038/nature08656

60. Timme RE, Allard MW, Luo Y, et al. Draft genome sequences of 21 Salmonella enterica serovar enteritidis strains. J Bacteriol 2012;194:5994–5. doi: 10.1128/JB.01289-12

61. Georgiades K, Raoult D. Defining pathogenic bacterial species in the genomic era. Front Microbiol 2010;1:151. doi: 10.3389/fmicb.2010.00151

62. Segerman B. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Front Cell Infect Microbiol 2012;2:116. doi: 10.3389/fcimb.2012.00116

63. Medini D, Donati C, Tettelin H, et al. The microbial pan-genome. Curr Opin Genet Dev 2005;15:589–94. doi: 10.1016/j.gde.2005.09.006

64. Caputo A, Merhej V, Georgiades K, et al. Pan-genomic analysis to redefine species and subspecies based on quantum discontinuous variation: the Klebsiella paradigm. Biol Direct 2015;10:55. doi: 10.1186/s13062-015-0085-2

65. Rouli L, Merhej V, Fournier P-E, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect 2015;7:72–85. doi: 10.1016/j.nmni.2015.06.005

66. Stewart EJ. Growing unculturable bacteria. J Bacteriol 2012;194:4151–60. doi: 10.1128/JB.00345-12

67. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet 2016;17:175–88. doi: 10.1038/nrg.2015.16

68. Borozan I, Watt S, Ferretti V. Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. Bioinformatics 2015;31:1396–404. doi: 10.1093/bioinformatics/btv006

3.2 *Integrating culture-based antibiotic resistance profiles with whole-genome sequencing data for 11,087 clinical isolates*

Genomics Proteomics Bioinformatics 17 (2019) 169–182

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Integrating Culture-based Antibiotic Resistance Profiles with Whole-genome Sequencing Data for 11,087 Clinical Isolates

Valentina Galata [1,a], Cédric C. Laczny [1,b], Christina Backes [1,c], Georg Hemmrich-Stanisak [2,d], Susanne Schmolke [3,e], Andre Franke [2,f], Eckart Meese [4,g], Mathias Herrmann [5,h], Lutz von Müller [5,i], Achim Plum [6,7,j], Rolf Müller [8,9,10,k], Cord Stähler [1,l], Andreas E. Posch [1,6,7,*,m], Andreas Keller [1,*,n]

[1] *Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany*
[2] *Institute of Clinical Molecular Biology, Christian-Albrechts University of Kiel, 24105 Kiel, Germany*
[3] *Siemens Healthcare GmbH, Strategy and Innovation, 91052 Erlangen, Germany*
[4] *Department of Human Genetics, Saarland University, 66421 Homburg, Germany*
[5] *Institute of Medical Microbiology and Hygiene, Saarland University, 66421 Homburg, Germany*
[6] *Ares Genetics GmbH, 1030 Vienna, Austria*
[7] *Curetis GmbH, 71088 Holzgerlingen, Germany*
[8] *Department of Pharmacy, Pharmaceutical Biotechnology, Saarland University, 66123 Saarbrücken, Germany*
[9] *Department of Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Saarland University, 66123 Saarbrücken, Germany*
[10] *Helmholtz Center for Infection Research and Pharmaceutical Biotechnology (HZI), Saarland University, 66123 Saarbrücken, Germany*

[*] Corresponding authors.
E-mail: andreas.keller@uni-saarland.de (Keller A), andreas.posch@ares-genetics.com (Posch AE).
[a] ORCID: 0000-0002-4541-427X.
[b] ORCID: 0000-0002-1100-1282.
[c] ORCID: 0000-0001-9330-9290.
[d] ORCID: 0000-0002-2896-4691.
[e] ORCID: 0000-0001-8409-411X.
[f] ORCID: 0000-0003-1530-5811.
[g] ORCID: 0000-0001-7569-819X.
[h] ORCID: 0000-0003-2638-2257.
[i] ORCID: 0000-0001-9013-4245.
[j] ORCID: 0000-0003-1635-1633.
[k] ORCID: 0000-0002-1042-5665.
[l] ORCID: 0000-0003-3453-0993.
[m] ORCID: 0000-0003-3893-3562.
[n] ORCID: 0000-0002-5361-0895.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2018.11.002

Received 4 June 2018; revised 9 October 2018; accepted 7 November 2018
Available online 14 May 2019

Handled by Yigang Tong

KEYWORDS

Antibiotic resistance;
Whole-genome sequencing;
Bacteria;
Pan-genome

Abstract Emerging **antibiotic resistance** is a major global health threat. The analysis of nucleic acid sequences linked to susceptibility phenotypes facilitates the study of genetic antibiotic resistance determinants to inform molecular diagnostics and drug development. We collected genetic data (11,087 newly-sequenced whole genomes) and culture-based resistance profiles (10,991 out of the 11,087 isolates comprehensively tested against 22 antibiotics in total) of clinical isolates including 18 main species spanning a time period of 30 years. Species and drug specific resistance patterns were observed including increased resistance rates for *Acinetobacter baumannii* to carbapenems and for *Escherichia coli* to fluoroquinolones. Species-level **pan-genomes** were constructed to reflect the genetic repertoire of the respective species, including conserved essential genes and known resistance factors. Integrating phenotypes and genotypes through species-level pan-genomes allowed to infer gene–drug resistance associations using statistical testing. The isolate collection and the analysis results have been integrated into GEAR-base, a resource available for academic research use free of charge at https://gear-base.com.

Introduction

The development of new antimicrobial drugs has largely stagnated over the last few decades [1], while the drug resistance rates of many pathogens have at the same time been increasing [2–4]. Various large-scale efforts have been launched to investigate the emerging drug resistance, such as the Meropenem Yearly Susceptibility Test Information Collection (MYSTIC) program [2], the Canadian National Intensive Care Unit (CAN-ICU) study [5], the Canadian National Surveillance (CANWARD) study [6,7], the Center for Disease Dynamics, Economics and Policy (CDDEP) study [3], and the European Antimicrobial Resistance Surveillance Network (EARS-Net) survey [8]. The results of these studies have shed light on the most common bacterial pathogens and resistance rates for regularly administered antibiotics, with the primary focus on the trend analysis of specific bacterial groups, periods of time, or locations [2,3,9–12]. The global challenge of emerging drug resistance is further exacerbated by the rising prevalence of microorganisms with multidrug resistance (MDR) phenotypes [13]. Accordingly, identifying and administering the most effective drug in each individual case is of even greater importance for successful treatment of bacterial infections. However, these studies did not investigate the genetic repertoire of the pathogens, which represents an important source of information—*e.g.*, the resistance genotype may be readily revealed while the respective phenotype is misleading or not expressed under artificial laboratory conditions [14,15].

Simultaneously, the recovery of genomic information from microorganisms via high-throughput sequencing approaches has become a routine task. This not only allows the high-resolution study of individual organisms' genomes, but also the aggregated study in the form of "pan-genomes"—the united genetic repertoire of a clade [16]. Pan-genomes can be used to identify common genetic potential—*i.e.*, the "core" genes of a clade—as well as genes that are less broadly conserved ("accessory" or "singleton" genes) [16]. This facilitates the identification of essential genes or genes that provide adaptation advantages. Multiple computational approaches are available for the systematic creation of pan-genomes, *e.g.*, Roary [17], EDGAR [18], and panX [19]. As a result, a variety of bacterial pan-genomes, typically at the species-level, have thus far been constructed [20–23]. However, most pan-genome studies focus on distinct species and do not always cover clinically relevant species. For example, MetaRef represents a resource that provides information about pan-genomes from multiple species and integrates approximately 2800 public genomes [24]. Although the diversity of the therein included organisms is particularly broad, the depth is limited in relation to clinically relevant bacteria—*e.g.*, seven *Klebsiella pneumoniae* genomes. Moreover, individual isolates included in the studies often span narrow time frames and/or have limited geographic spread.

While pan-genomic studies typically focus on the genetic information alone, efforts combining genomic and phenotypic information, in particular from antibiotic resistance testing, for the study of conserved or emerging resistance mechanisms are becoming increasingly prevalent [25–28]. There are many antibiotic resistance resources available [29], however only few link genomic and phenotypic information of bacterial isolates. One of such resources is the Pathosystems Resource Integration Center (PATRIC) [30], which represents a rich service for the study of >80,000 genomes [31]. Yet, antimicrobial resistance information is available only for about 10% of the genomes. Furthermore, as the genomes and the associated metadata of PATRIC are imported from public resources, which are populated by individual research efforts, data standardization or normalization is challenging. Finally, individual taxa may be underrepresented and thus warrant expansion—*e.g.*, the number of *Escherichia spp*. genomes with antimicrobial resistance metadata is almost two orders of magnitude smaller than that of *Mycobacterium spp*. genomes [31].

Motivated by the importance of linking resistance phenotypes with genomic features, we collected whole-genome sequencing data of 11,087 clinical isolates representing, *inter alia*, 18 main bacterial species. The samples were collected in North America, Europe, Japan, and Australia over a period of 30 years, and processed in a concerted effort, thereby reducing experimental bias. Culture-based resistance testing was performed for 10,991 out of the 11,087 isolates against 22 antibiotic drugs. Furthermore, species-level pan-genomes were constructed on the basis of per-isolate *de novo* assemblies and were used to infer gene–drug resistance associations. This wealth of information is integrated into an online resource, Genetic Antibiotic Resistance resource, or in short, GEAR-base (**Figure 1**). Providing broad organismal, antibiotic treatment and temporal coverage, GEAR-base is expected to support the pan-genome-based study of bacteria and to advance research on known or emerging antibiotic resistance mechanisms. GEAR-base is available for academic research use free of charge at https://gear-base.com.

Results

Resistance testing of cultured bacterial isolates

The present dataset of 11,087 bacterial isolates covered a total of 6 families, 14 genera, and 20 species (considering species with at least 50 isolates, Table S1) and comprised two datasets: 1001 isolates from the *Staphylococcus aureus* strain collection and 10,086 isolates from the Gram-negative collection. From the *S. aureus* strain collection, 993 isolates were tested for methicillin resistance and susceptibility (see Methods section). For 9998 isolates from the Gram-negative collection, culture-based antimicrobial susceptibility testing (AST) for 21 commonly-prescribed Food and Drug Administration (FDA)-approved antibiotics from 8 drug classes was performed to determine the respective minimum inhibitory concentrations (MICs) (**Figure 2**A). The resistance profiles were determined for each isolate in accordance with the European Committee on Antimicrobial Susceptibility Testing (EUCAST) guidelines (v. 4.0) for a total of 182 drug concentrations (7–11 concentrations per drug; Tables S2 and S3, Figure 2B). Whole-genome sequencing (WGS)-based taxonomic identification was performed for all isolates [32]. In the following content, we focused on the analysis results of the MICs and resistance profiles of the 9998 isolates from the Gram-negative collection.

All patient-derived isolates were collected in clinics located in North America, Europe, Japan, or Australia from 1983 to 2013 (Figure S1). Varying degrees of resistance were observed among the isolates (Figure 2B). The majority of species demonstrated relatively low resistance rates (< 20%) to aminoglycosides (gentamicin and tobramycin) and carbapenems (ertapenem, imipenem, and meropenem), except for *Acinetobacter baumannii* (≥ 29% for aminoglycosides and meropenem), *Pseudomonas aeruginosa* (26% for gentamicin), and *Klebsiella pneumoniae* (26% for tobramycin). These rates were compared against two independent large-scale studies—CDDEP (USA-based results; CDDEP ResistanceMap, https://resistancemap.cddep.org/AntibioticResistance.php, accessed on September 26, 2017) [3] program and the MYSTIC program [2], for matching species and drug data. Both studies report low (< 20%) resistance rates for the aminoglycosides and carbapenems during the observation period (1999–2012/2014 for CDDEP and 1999–2007/2008 for MYSTIC) except for *A. baumannii* (CDDEP: > 20% since 2005 for carbapenems and > 35% during 1999–2012 for aminoglycosides; MYSTIC: > 37% in 2007/2008 for carbapenems and > 20% during most years for aminoglycosides). For *K. pneumoniae* and tobramycin (aminoglycosides for CDDEP), MYSTIC and CDDEP reported > 10% resistance rates since 2005 with only one value of above 20% observed by MYSTIC in 2007. Finally, for *P. aeruginosa* and gentamicin, MYSTIC reported a resistance rate of only around 10%. The rate of isolates resistant to multiple antibiotic drugs, *i.e.*, resistant to at least three drugs from different drug classes (CDDEP ResistanceMap), was highest for *A. baumannii* (44%) and for *Enterobacter* spp. (41%–45%). For the remaining species and drug classes,

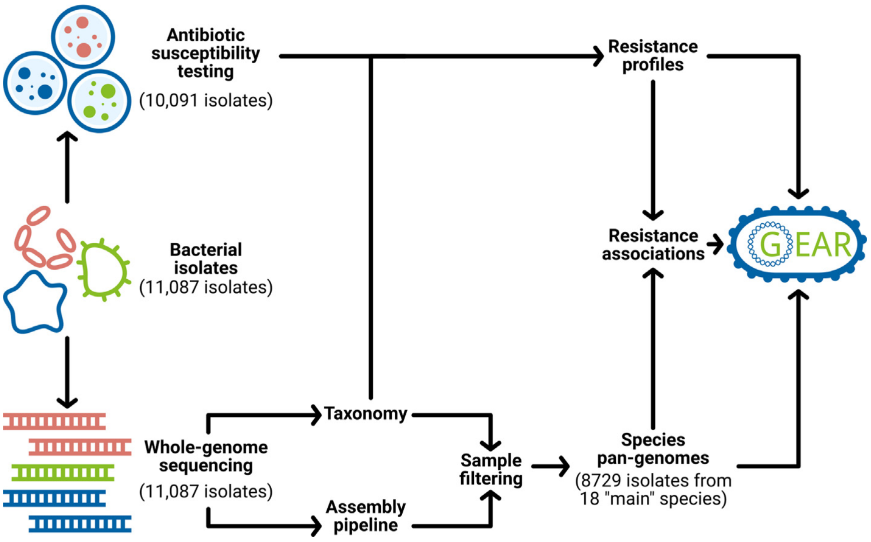


Figure 1 GEAR-base workflow and structure
Schematic overview of data collection, processing and integration into GEAR-base.

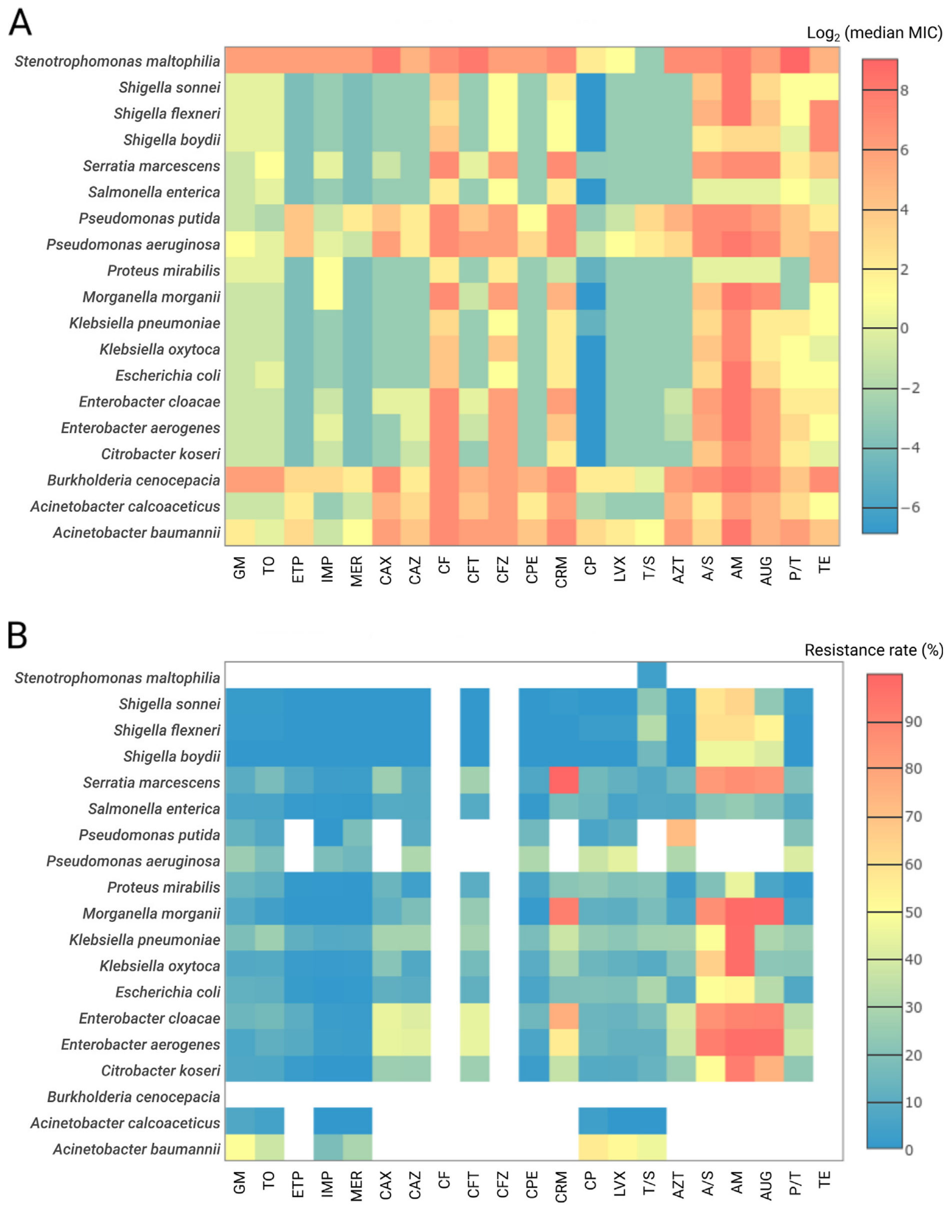


Figure 2 Overview of resistance profiles

Heatmaps of log-transformed (base 2) median MIC values (**A**) and resistance rates (**B**) for all species with at least 50 isolates. Drugs labels were grouped relative to their class. The cells are coded in color gradient from blue to red with blue for lower values and red for higher values. White color in panel B corresponds to the cases where no breakpoints are available from the used guidelines. MIC, minimum inhibitory concentration.

the MDR rates were at least 20%, except for *Acinetobacter calcoaceticus* (0%), *Salmonella enterica* (11%), and *Shigella* spp. (0%–3%). In addition to the investigation of individual species–drug combinations, we analyzed whether drug pairs showed correlating MIC profiles among all isolates (Figure S2). In general, the highest correlations were expectedly found within separate drug classes — *e.g.*, for fluoroquinolones, aminoglycosides, and carbapenems. While for some species, *e.g.*, *Burkholderia cenocepacia*, a clear clustering according to drug classes and their mechanism of action was observed, other species, such as *S. enterica*, showed less pronounced cluster structures.

Subsequently, we compared resistant and non-resistant isolates with respect to their collection year in order to identify potential trends of de-/increasing antibiotic resistance rates (Figures S3 and S4, and Table S4). The following species–drug pairs were found to exhibit particularly low *P* values [WMW-test, false discovery rate (FDR) adjusted $P < 1E-17$], as well as increases in resistance over time: *K. pneumoniae* to cefepime, *K. pneumoniae* and *A. baumannii* to carbapenems, and *E. coli* to fluoroquinolones. Similar trends were reported by the CDDEP [3] program (CDDEP ResistanceMap) and the MYSTIC program [2], including increasing resistance rates for *A. baumannii* to carbapenems (43% from 1999 to 2014 in the USA, CDEEP), and for *E. coli* to fluoroquinolones (30% from 1999 to 2014 in the USA, CDEEP; >20% from 1999 to 2008, MYSTIC).

While the culture-based analyses provide species-resolved information about resistance rates over time and corroborate previous findings on the global increase in antibiotic resistance, genetic features represent important factors and were thus concomitantly considered.

Whole-genome *de novo* assembly of isolates and species pan-genomes

A total of 11,087 bacterial isolates were whole-genome sequenced using Illumina Hiseq2000/2500 sequencers, resulting in a median number of 1,517,147 paired reads per isolate (1,609,533 ± 620,481). *De novo* assemblies were successfully created for 11,062 (99.8%) isolates (**Figure 3**) and of these, the assembled genomes of 10,764 (97.3%) isolates passed the stringent assembly quality criteria. Moreover, the assembled genomes of 9206 (83% of 11,087) isolates fulfilled the quality criteria for taxonomic assignment. A total of 8729 isolates, representing 18 main species having ≥50 isolates, were used after stringent quality filtering (see Methods for sample filtering details) in the subsequent analyses and in the construction of species-level pan-genomes (Table S3).

First, the presence/frequency of genes from a set of 111 single-copy marker genes, which were defined as essential genes by Dupont et al. [33], was used as a proxy to estimate the genome completeness of individual *de novo* assemblies. Overall, the assemblies were found to be largely complete. 92 essential genes (82.9%) were identified in at least 99% of the

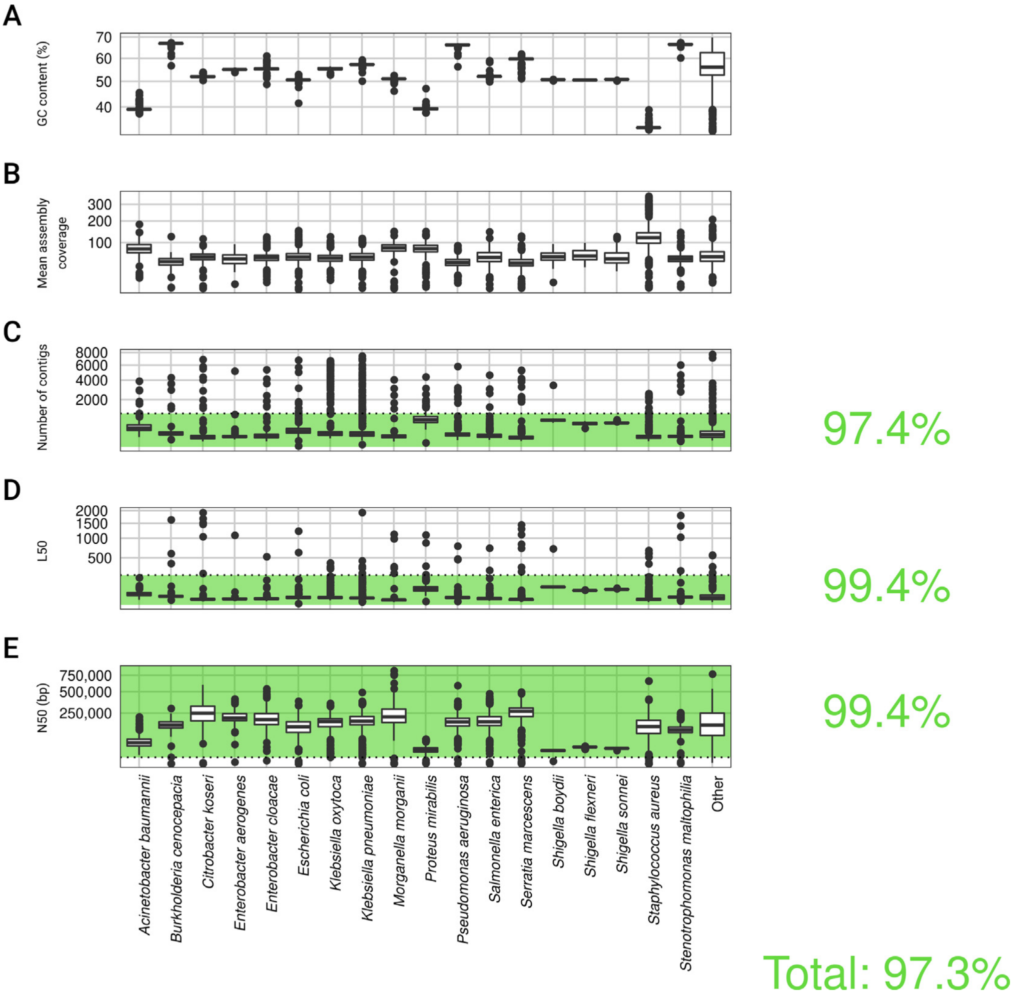


Figure 3 Assembly quality overview

Assembly summary statistics for the 11,062 isolates with a *de novo* assembly. The isolates were grouped by their species taxon, and isolates not belonging to any of the main 18 species used for pan-genome construction were grouped into ”Other”. The box plots show the GC content (**A**), mean assembly coverage (**B**), number of contigs (**C**), L50 value (**D**), and N50 value (**E**) for contigs of at least 200 bp. The assembly quality cut off values are illustrated by dotted lines (1000 for the number of contigs; 200 for L50; and 5000 bp for N50). The plot area satisfying the respective filtering criterion is colored in green. Percentages of isolates passing the respective criterion as well as all criteria are shown to the right.

8729 isolates (Figure S5) that were used to construct a phylogenetic tree of these isolates (Figure S6). Furthermore, species-specific presence/absence patterns were frequently observed (Figure S7A). For example, TIGR00389 (glycine–tRNA ligase) was only found in *S. aureus*, whereas TIGR00388 (glycine–tRNA ligase, alpha subunit) was not present in this species. Four genes, TIGR00408 (encoding the proline–tRNA ligase), TIGR02387 (encoding the DNA-directed RNA polymerase, gamma subunit), TIGR00471 (encoding the phenylalanine–tRNA ligase, beta subunit), and TIGR00775 (encoding the Na^+/H^+ antiporter, NhaD family), were not found in any of the isolates, except for sporadic hits in *Pseudomonas aeruginosa* for TIGR00408.

In the next step, Resfams core-based resistance factors [34] were annotated in the isolate assemblies in order to study the species-level distribution of these genetic features. The number of covered Resfams (mean count of hits ≥1) varied between species from 4.1% (5 of 123 Resfams, *Morganella morganii*) to 11.4% (14 of 123 Resfams, *A. baumannii* and *Shigella sonnei*) (Figure S8). Three Resfams were found in at least 90% of all considered isolates. These are all antibiotic efflux pumps, which include RF0007 [ATP-binding cassette (ABC) type], RF0107 (ABC type), and RF0115 [resistance-nodulation-cell division (RND) type], with the latter having a mean count of hits of ≥5 for 14 out of 18 species.

The multi-locus sequence typing (MLST) analysis revealed, that in all species with a typing scheme included in the used version of PubMLST, isolates were assigned to at least 6 different sequence types (STs), except for *S. sonnei*, and new STs could be identified, except for *Shigella flexneri* and *S. sonnei* (Figure S9). Among these species, the proportion of isolates without a confident assignment was high (≥10%) for *B. cenocepacia*, *Enterobacter cloacae*, *Klebsiella oxytoca*, and *Stenotrophomonas maltophilia*.

The size of the species pan-genomes (*i.e.*, the number of centroids) ranged from 5838 (*S. aureus*, total pan-genome length < 5 Mb) to 42,046 (*E. cloacae*, total pan-genome length > 30 Mb) (Figure S10). A centroid refers here to the representative gene of a homologous gene cluster with ≥90% pair-wise amino acid sequence identity (Methods). Most centroids were found in < 10% or in ≥90% of the isolates (**Figure 4**). Moreover, all pan-genomes were found to be open based on the analysis of the number of centroids in relation to the number of included genomes (Figure S11, Table S5). The two-dimensional embedding of the core centroids from the pan-genomes revealed many taxon-specific patterns (Figure S12) with distinct clusters for *B. cenocepacia*, *M. morganii*, *A. baumannii*, *Proteus mirabilis*, *S. aureus*, *S. maltophilia*, *P. aeruginosa*, and *Serratia marcescens*. We compared the number of (core) centroids in our pan-genomes to the numbers reported by panX [19] (http://pangenome.tuebingen.mpg.de, accessed on January 29, 2018). The number of centroids present in at least 90% of the analyzed genomes was consistent for all matching species (Table S6). However, the pan-genome size, *i.e.*, the total number of centroids described in GEAR-Base, was similar for *E. coli* and *S. aureus*, but exceeded substantially the number of centroids described in panX for *A. baumannii*, *K. pneumoniae*, *P. aeruginosa*, and *S. enterica* (Table S6). With respect to the presence of essential genes in the species-level pan-genomes, the mean number of centroids containing at least one matching gene was one, that is, these essential genes were mostly found in only one centroid cluster (Figure S7B). However, the mean number of centroids was ≥1.25 for eight essential genes, *i.e.*, in some species these genes were found in multiple centroid clusters.

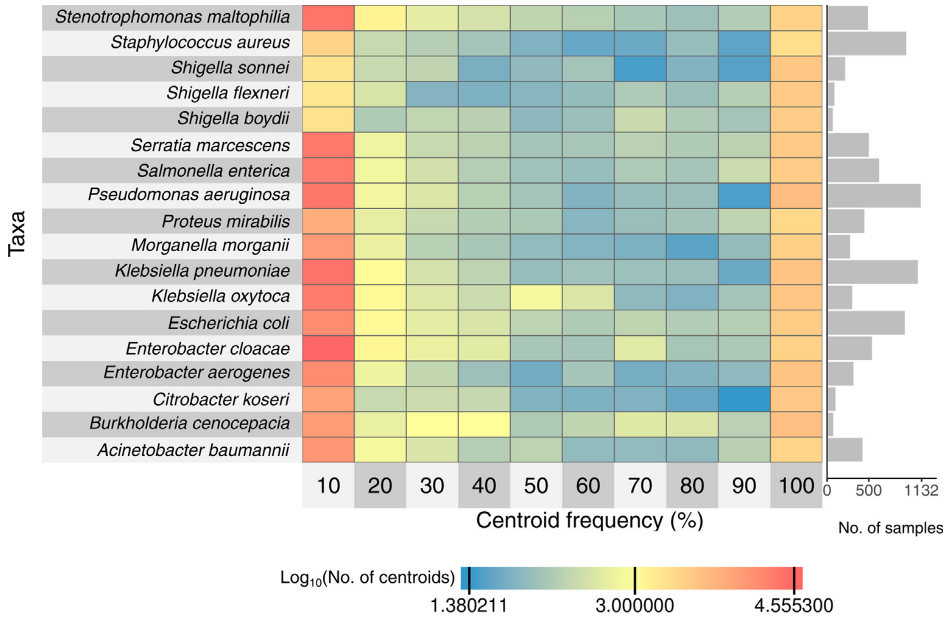


Figure 4 Centroid frequency

Number of centroids in each pan-genome of the 18 main species in relation to their frequency. The first column contains centroids that are present in < 10% of the isolates, and the last one contains centroids that are present in ≥90% of the isolates. Cells are coded in color gradient to indicate the log10-transformed number of centroids. The bar plot on the right shows the number of isolates used to construct the respective pan-genomes.

In the following section, the resistance phenotypes and genomic features were linked and significantly associated centroids were further studied, with respect to their overlap to known resistance genes from the Resfams core database.

Resistance associations by linking phenotype and genotype

We used binary information in the form of centroid presence/absence to test for significant centroid–drug associations per species. The number of found associations ranged from below 10 to above 500; most associations (≥500) were found for *P. aeruginosa* and tobramycin, and *K. pneumoniae* and gentamicin (**Figure 5**). Furthermore, the drug resistance-associated centroids encoding for a resistance gene were investigated. From the Resfams core database, 45 of the 123 factors were found in at least one centroid (Figure S13). Among these, the top ten Resfams genes from both analyses covered various resistance mechanism classes – nucleotidyltransferases, phosphotransferases, acetyltransferases, beta-lactamases, and major facilitator superfamily (MFS) transporters (Figure S13B).

GEAR-base online resource

The GEAR-base resource is freely accessible at https://gear-base.com for academic research use and currently provides two modules for browsing of the database—a culture-based module and a pan-genome module—as well as a module for the analysis of user-provided data. The culture-based module is focused on the Gram-negative isolate collection and provides an interactive view of the taxonomic composition, MIC, and resistance profiles, as well as additional meta-data, *e.g.*, collection year or isolate distributions. The pan-genome module provides general statistics, such as assembly quality of the included isolates, pan-genome size, and resistance association analysis overview, for both the Gram-negative and the *S. aureus* isolates. Gene nucleotide sequences can be downloaded for each individual pan-genome centroid and a batch-download of all centroid nucleotide sequences is available. Moreover, pan-genome centroids can be browsed online for specific gene products and filtered by their presence in the isolates. In addition, centroid clusters can be viewed including associated gene annotations, the hits to the Resfams core database, and information about potential resistance associations against the set of herein included drugs. GEAR-base's analysis module allows the user to query individual gene sequences against the pan-genome centroid sequences using Sourmash [35], against hidden Markov models (HMMs) of pan-genome centroid clusters and Resfams core database using HMMER (http://hmmer.org/), and against the NCBI nt/nr database using BLASTp [36]. Furthermore, a genome-scale search against the present clinical isolate collection, the finished genomes from the NCBI RefSeq database [37], as well as the National Collection of Type Cultures (NCTC) 3000 genomes project from the Public Health England and the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/resources/downloads/bacteria/nctc/, accessed on October 18, 2017) can be performed online using Mash/MinHash [38].

We used a recently-published *K. pneumoniae* genome [39] (strain 1756, NCBI assembly accession ID GCF_001952835.1_ASM195283v1) to demonstrate the analysis functionalities of GEAR-base. In a first step, the chromosome and plasmid sequences were uploaded and a perfect match was found to the genome's NCBI entry, as expected. The next-best matches were to a *K. pneumoniae* isolate from the current collection of clinical isolates (828/1000 shared hashes, distance of 4.71E−3), and to a *Klebsiella* sp. genome (ERS706555) from the NCTC 3000 database (709/1000 shared hashes, distance of 8.89 E-3). In a second step, all coding DNA sequences (CDS) were searched against the pan-genome centroids in GEAR-base using Sourmash and against the Resfams core database.

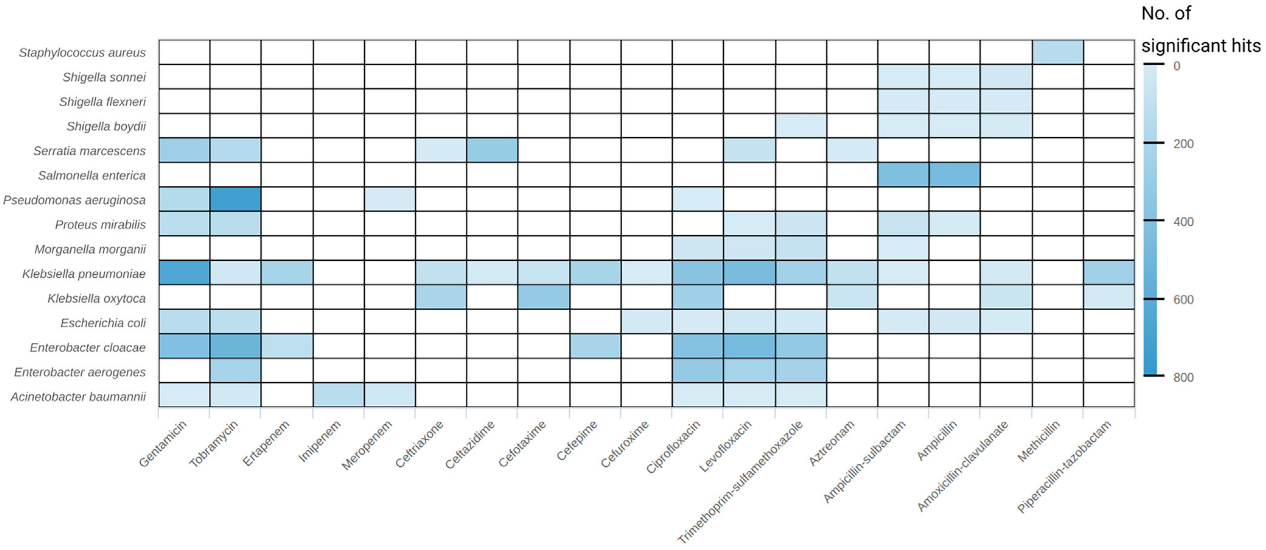


Figure 5 Number of significant results of the resistance association analysis
Significant results (adjusted $P < 1E-5$) of the resistance association analysis based on the presence/absence of centroids. The heatmap shows the number of significant results (in color gradient with lighter blue for smaller numbers and darker blue for larger numbers) per taxon and drug. Drugs are sorted according to their class.

The majority of the pan-genome hits were related to *K. pneumoniae* (6206 hits of 11,267) followed by *E. aerogenes* (1537 hits) and *K. oxytoca* (1014 hits). *S. aureus*, a Gram-positive species, served as an outgroup and no hits to its pan-genome were found. In total, 37 hits to 21 unique Resfams (core database) were found in the query genome CDS with 23 hits on the chromosome and 14 on the plasmid. The top three most occurring Resfams were RF0115 (8 hits, RND antibiotic efflux pump), RF0098 (3 hits, multidrug efflux RND membrane fusion protein MexE, RND antibiotic efflux), and RF0053 (3 hits, class A beta-lactamase). Furthermore, the CDSs of eight antibiotic resistance genes reported in the original genome announcement were investigated. The HMM-based search of pan-genome centroids resulted in the identification of two chromosomal CDSs, WP_076027158.1 (multidrug efflux RND transporter periplasmic adaptor subunit OqxA) and WP_004146118.1 (FosA family fosfomycin resistance glutathione transferase), being classified as *K. pneumoniae*-derived centroids according to their top hits (with respect to the full sequence score). The top hits of the remaining genes (5 plasmid-derived and 1 chromosome-derived) included centroids from other Gram-negative species. However, the centroid cluster annotations matched the expected protein functions for all eight CDSs independent of the species. The top three hits for WP_004146118.1 were centroids from *K. pneumoniae*, *E. aerogenes*, and *K. oxytoca*, matching the expected annotation and present in almost all isolates (>98%) of the respective pan-genomes. This high prevalence matches the observations made by Ryota *et al.* reporting similarly high frequency (>96%) of *fosA* in these species [40]. For the beta-lactamases WP_004176269.1 (class A broad-spectrum beta-lactamase SHV-11) and WP_000027057.1 (class A broad-spectrum beta-lactamase TEM-1), the top hits in *Klebsiella* were associated with resistance to penicillins and cephalosporins. And for the aminoglycoside transferases WP_000018329.1 (aminoglycoside *O*-phosphotransferase APH(3′)-Ia), WP_032491824.1 (ANT(3″)-Ia family aminoglycoside nucleotidyltransferase AadA22), and WP_000557454.1 (aminoglycoside *N*-acetyltransferase AAC(3)-IId), the top hits in *K. pneumoniae* were associated with resistance to aminoglycosides. Moreover, all three chromosome-derived CDSs (WP_004176269.1, WP_076027158.1, and WP_004146118.1) matched to centroids found in >92% of the *K. pneumoniae* isolates, two of the five plasmid-derived CDSs (WP_032491824.1 and WP_000027057.1) matched to centroids with a frequency of >25%, while the remaining CDSs matched to centroids with a frequency of <12%.

Discussion

To facilitate the studies on antibiotic resistance, we have built GEAR-base, a resource incorporating paired data on resistance phenotypes and genomic features for an extensive, longitudinal collection of clinical isolates from various bacterial species. This concerted effort is expected to reduce experimental bias and the present resource provides a portal for information retrieval as well as data analysis.

Species-level antibiotic resistance phenotypes can be inspected using the culture-based module in GEAR-base. Specifically, resistance rates and trends across multiple species and antibiotic drugs can be assessed on a large scale, which we believe is important for current and future antibiotic resistance research. Although some effect of potential sampling bias cannot be excluded, our findings on the increased resistance rates corroborate previously reported trends. In addition to this phenotypic information, genomic information is included in the pan-genome module. Such information can be used independent of the phenotypic information, *i.e.*, purely from a pan-genomic perspective, *e.g.*, for the study of inter- or intra-species gene conservation. The observed number of core centroids was consistent with the statistics reported by panX. However, GEAR-base pan-genomes are based on significantly higher sample number and are substantially larger in size, thus giving access to a comprehensive collection of the genome heterogeneity for human bacterial pathogens. In addition, GEAR-base links these two information layers through centroid–drug associations. These associations can subsequently be explored to study resistance mechanisms. Furthermore, individual researchers can compare genes or genomes of interest to the present resource, thereby providing an independent layer of support. This functionality was demonstrated using a recently published carbapenem-resistant *K. pneumoniae* isolate. While the taxonomic classifications of the genome and of a set of chromosome-derived antibiotic resistance genes are consistent with the expected taxonomy of the isolate, the plasmid-derived antibiotic resistance genes exhibit ambiguous taxonomic assignments, which is not unexpected for plasmid-borne genes. Moreover, the extensive collection of isolates included herein enables the study of the overall conservation degrees and the time-resolved frequencies of this exemplary antibiotic resistance gene set.

The analysis functionality in GEAR-base covers external genome databases (NCBI RefSeq as well as the NCTC 3000 genomes project from Public Health England and the Wellcome Trust Sanger Institute) in addition to the present collection of clinical isolate genomes. However, because the majority of external genomes are not linked to antibiotic resistance information and centroid–drug associations are considered a key component of the present resource, the pan-genome module is restricted to the present isolates. Additionally, the species-level pan-genome centroids in GEAR-base are available for download and provide a great opportunity for subsequent integration with external genomes for further study.

Emerging antibiotic resistance represents a multidisciplinary and global challenge. We believe that GEAR-base will serve as a valuable resource enabling the detailed analysis of resistance-associated genomic features. GEAR-base includes a comprehensive selection of clinically highly relevant human microbial pathogens and will thus be of great use for the research and clinical communities.

Materials and methods

Bacterial isolates

The dataset of 11,087 isolates consisted of 1001 isolates from the *S. aureus* strain collection of Saarland University Medical Center and a collection of 10,086 Gram-negative bacterial clinical isolates that form part of the microbiology strain collection of Siemens Healthcare Diagnostics (West Sacramento, CA) [32]. DNA extraction using the Siemens VERSANT®

sample preparation system [41] and whole-genome next-generation sequencing were performed for all isolates as described in Galata et al. [32] (2 × 100 bp paired-end on Illumina Hiseq2000/2500 sequencers).

Methicillin susceptibility of *S. aureus* isolates

For 993 isolates from the *S. aureus* strain collection, detection of methicillin-resistant or methicillin-susceptible *Staphylococcus aureus* (MRSA/MSSA) isolates was performed. The specimen were plated on CHROMagar MRSA detection biplates (Mast, Reinfeld, Germany). All MRSA-positive culture isolates were further confirmed using a penicillin-binding protein 2a latex agglutination test (Alere, Köln, Germany).

Susceptibility testing and resistance profiles of Gram-negative isolates

For 9998 isolates from the Gram-negative isolate collection, AST was performed. Frozen reference AST panels were prepared following Clinical Laboratory Standards Institute (CLSI) recommendations [42]. The antimicrobial agents included in the panels are provided in Table S2. Prior to use with clinical isolates, AST panels were tested and considered acceptable for testing with clinical isolates when the QC results met QC ranges described by CLSI [42].

Isolates were cultured on trypticase soy agar with 5% sheep blood (Bethesda Biological Laboratories, Cockeysville, MD) and incubated in ambient air at 35 ± 1 °C for 18–24 h. Isolated colony panels were inoculated according to the CLSI recommendations (CLSI additional reference) and incubated in ambient air at 35 ± 1 °C for 16–20 h. Panel results were read visually, and MICs were determined.

MIC value processing

The bacterial culture may not grow for the lowest drug concentration tested (expressed as $\leq x$) or show no significant growth decrease for the highest concentration tested (expressed as $>x$), where x represents the drug concentration tested. To allow consistent processing, these MIC values were transformed as follows: in the former case, the MIC value was set to be $x/2$ (*e.g.*, "≤0.25" was set to "0.125"), and in the latter case, the MIC value was set to be $x * 2$ (*e.g.*, ">64" was set to "128"). Additionally, we considered only the MIC value of the first agent in case of drug combinations (*e.g.*, "32/16" was set to "32").

Drug information

The 21 drugs used in this study were grouped into 8 drug classes based on their category in the EUCAST guidelines [43]. Among them, 7 drugs belong to cephalosporins (cefazolin and cephalotin – 1st generation; cefuroxime – 2nd generation; cefotaxime, ceftazidime, and ceftriaxone – 3rd generation; and cefepime – 4th generation), 4 to penicillins, 3 to carbapenems, 2 to fluoroquinolones, 2 to aminoglycosides, and 1 to tetracycline. In addition, 1 drug is a monobactam and the remaining 1 drug falls into the category "miscellaneous" (Table S2).

Resistance classification

EUCAST guidelines [43] (v. 4.0) were used for MIC value classification. Isolates were classified as resistant, intermediate, or susceptible. An isolate was considered to be resistant if the corresponding MIC value was greater than the resistance breakpoint. If the MIC value was below or equal to the susceptibility breakpoint, the isolate was considered to be susceptible. If the MIC value was between the two breakpoints, the isolate was considered as "intermediate". If no breakpoint was available for a specific drug and bacterial group, no classification was performed.

Genome-based taxonomic classification

Kraken [44] (v. 0.10.4-beta) was used with the default database containing finished genomes from the NCBI RefSeq database (accessed on January 13, 2015) and a k-mer length of 31. Report files were created from the raw output using "kraken-report" and processed to retrieve the information, including (1) the first best species hit relative to the percentage of mapped sequences; (2) the number of sequences mapped to best hit; (3) the number of sequences classified at species level; (4) the number of unclassified sequences; and (5) the total number of reported sequences. In addition, sensitivity values, precision values, and percentages of unassigned sequences were calculated. Sensitivity was defined as the ratio of reads assigned to the best hit over the total number of reported reads. Precision was defined as the ratio of reads assigned to the best hit over reads classified at species level. For each sample, the taxonomic lineage from the species to the class level was retrieved using the R package "taxize" [45] and the NCBI [46] taxonomy database (accessed on February 8, 2016). An overview of the taxonomic composition of the dataset was created using Krona [47].

Read processing and assembly pipeline

The raw sequencing reads were trimmed using Trimmomatic [48] (v. 0.35, command line parameters: PE ILLUMINACLIP:NexteraPE-PE.fa:1:50:30 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). Trimmed paired-end reads were assembled *de novo* into scaffolds (from now on called contigs for simplicity) using SPAdes [49] (v. 3.6.2, parameters: -k 21,33,55 --careful) and annotated by Prokka [50] (v. 1.11, parameters: --gram neg --mincontiglength 200). Assembly quality was assessed using QUAST [51] (v. 3.2, parameters: --contig-thresholds 0,100,200,500,1000 --min-contig 200).

Mean assembly coverage

Trimmed reads were mapped to the contigs (minimal length of 200 bp) using BWA [52] (v. 0.7.12) and SAMtools [53] (v. 1.2; command line: bwa mem –M –t <cores> <contigs> <forward reads> <reverse reads> | samtools view @ <cores> -bt <contigs> - | samtools sort -@ <cores> - <bam>). Then coverage histogram was computed using BEDtools [54] (v. 2.25; parameters: bedtools genomecov –ibam <bam> -g <contigs> > <hist>). Finally mean coverage was computed over all contigs.

Essential genes

Essential genes as defined by Dupont et al. [33] were downloaded (https://github.com/MadsAlbertsen/multi-metagenome/raw/master/R.data.generation/essential.hmm, accessed on March 7, 2017) and searched in the present assemblies (protein FASTA files of translated CDS; *.faa) using hmmsearch from the HMMER software package (http://hmmer.org/, v. 3.1b2, parameters: --cut-tc). Only hits with at least one domain satisfying the reporting thresholds (column "rep" in table output files) were considered. Best hits for each isolate and essential gene were determined with respect to the E-value of reported full sequences. Finally, each considered hit was assigned to a centroid, *i.e.*, the centroid covering the gene from the corresponding hit.

Resistance factors

The Resfams core database [34] of HMMs (v1.2) was used to identify known resistance factors in the present assemblies (*.faa, FASTA file of protein annotations) using hmmsearch from the HMMER software package (http://hmmer.org/, v. 3.1b2, parameters: hmmsearch --cut_ga --tblout output.tblout Resfams.hmm input.faa > output.hmmout).

MLST profiles were determined using the BLASTn search-based tool mlst (https://github.com/tseemann/mlst, accessed on August 8, 2016, v. 2.9, parameters: --minid 99 --mincov 75 --minscore 99) on assembled contigs (minimal length of 200 bp).

Sample filtering

First, the bacterial isolate samples were filtered on the basis of their taxonomic assignment and assembly quality. For the taxonomic assignments, the minimal sensitivity was set to 50% (0% for *Shigella*), the minimal precision to 75% (60% for *Shigella*), and the minimal percentage of unclassified reads to 30%. The cutoff values were "relaxed" for *Shigella* because of the well-known problem of high genetic similarity between the *Shigella* species and *E. coli* [55], making it difficult to differentiate between these organisms at the nucleotide level, which affects the taxonomic sensitivity. For the *de novo* assemblies, we used the criteria defined by RefSeq [37]: number of contigs $\leq$1000, N50 $\geq$5000, and L50 $\leq$200. Isolates that passed both filtering steps were grouped by their species taxon, and only species containing at least 50 isolates were further considered. As a result, the following 18 species (referred to as "main species" in the manuscript) passed the filtering step. These include *A. baumannii*, *B. cenocepacia*, *Citrobacter koseri*, *E. aerogenes*, *E. cloacae*, *E. coli*, *K. oxytoca*, *K. pneumoniae*, *M. morganii*, *P. mirabilis*, *P. aeruginosa*, *S. enterica*, *S. marcescens*, *Shigella boydii*, *S. flexneri*, *S. sonnei*, *S. aureus*, and *S. maltophilia*. Additionally, samples containing more than 10 essential genes in multiple copies were examined further by running Kraken (k = 31) on the nucleotide sequences of the annotated genomic features (*.ffn). Report files were created from filtered assignments (kraken-filter, threshold 0.05) and inspected manually in order to determine whether a large percentage of sequences was assigned to unexpected species. In total, 8729 isolates remained assigned to the 18 main species mentioned above.

Pan-genome construction

Roary [17] (v. 3.5.7, parameters: -e -n -i 90 -cd 90 -a -g 70,000 -r -s -t 11) was used to construct the species-level pan-genomes.

Centroid HMMs

The protein sequences were extracted from the FASTA files of translated CDS (*.faa) created by Prokka [50]. For non-CDS sequences, protein sequences were created by translating the corresponding nucleotide sequences from the nucleotide FASTA files (*.ffn) using BioPython (parameters: table = 11, stop_symbol = "*", to_stop = False, cds = False). Multiple sequence alignments were created using MUSCLE [56] (v. 3.8.31, parameters: -maxiters 1 -diags -sv -distance1 kbit20_3). HMM profiles were calculated using hmmbuild from the HMMER software package (http://hmmer.org/, v. 3.1b2).

Database

The GEAR-base was implemented using the Python web framework Django (v. 1.9.5) and MySQL (v. 15.11) as the database management system. HMM search in Resfams core database and centroid HMM profiles is implemented using package/library HMMER (http://hmmer.org/, v. 3.1b1). Moreover, sketches of centroid nucleotide sequences were computed using Sourmash [35] (v. 2.0.0.a1, sketching parameters: sourmash compute --dna --singleton --scaled 10 --seed 42 --ksizes 21, indexing parameters: sourmash index --dna --ksize 21). Mash/MinHash [38] (v. 1.1.1, default parameters) was used to create sketches of GEAR-base isolates, finished bacterial genomes from the NCBI RefSeq database, and assembled bacterial genomes from the NCTC 3000 database of Public Health England and the Wellcome Trust Sanger Institute. The genomes from the NCBI RefSeq database included 7118 genomes and were downloaded on June 17, 2017 using the NCBI genome downloading scripts of Kai Blin (https://github.com/kblin/ncbi-genome-download, accessed on October 18, 2017, v. 0.2.2) with the setting "ncbi-genome-download --section refseq --assembly-level complete --human-readable --parallel 10 --retries 3 --verbose bacteria" with "--format fasta" and "--format cds-fasta"). The bacterial genomes from the NCTC 3000 database were downloaded on July 10, 2017 and included 1052 genomes.

Resistance profile analysis of cultured isolates from the Gram-negative collection

Drug correlations

Considering only species with $\geq$50 isolates, pairwise drug correlations were computed using the MIC value profiles (Spearman's correlation coefficient, all isolates and for each species taxon separately). Drugs with a single MIC value across all considered isolates were removed prior to correlation computation. To visualize possible drug–drug associations, hierarchical clustering using Euclidean distance and average linkage was applied.

Association between isolate collection year and resistance profiles

Two-sided WMW-test (R package exactRankTests, v. 0.8-29) was applied to the isolates with assigned collection year available and belonging to a species taxon with ≥50 isolates (in total 8768 isolates from 18 taxa). The isolates were divided into resistant and non-resistant (susceptible and intermediate) groups. No test was performed if either group included < 10 isolates or all isolates in a group were collected in the same year. All *P* values were adjusted using FDR.

Phylogenetic analysis

Essential genes, found in ≥99% of the isolates that were used to construct the pan-genomes, were identified. Protein sequences for the corresponding best hits were extracted for each essential gene and isolate. Multiple sequence alignments were computed using MUSCLE [56] (v. 3.8.31, parameters: -maxiters 1 -diags -sv -distance1 kbit20_3) for each essential gene separately and concatenated into one alignment. If an isolate did not have any matches, an empty alignment sequence (*i.e.*, containing only gap characters) was added. RAxML [57] (v. 8.2.9, raxmlHPC-PTHREADS) was used to construct a phylogenetic tree from the aggregated alignment. After removing sequence duplicates (2297 in total) and alignment columns containing only undetermined values, *i.e*. ambiguous characters, (147 in total), the tree was built using the CAT model (parameters: -p 12,345 -m PROTCATAUTO -F -T 30).

Pan-genome analysis

Centroid rate estimation

The centroid presence–absence tables created by Roary were used to estimate the median number of total, new, unique, and core centroids in species-level pan-genomes relative to the number of isolates used (rarefaction). For each pan-genome, the columns (isolates) of the table were permuted 100 times. Starting from the first isolate, centroid counts were calculated in a cumulative manner for each permutation. The centroid categories were defined as follows: total centroids comprise centroids found in at least one of the included genomes; new centroids refer to the centroids found only in the last included genome; unique centroids are centroids found only in one of the included genomes; and core centroids are centroids found in ≥90%, ≥95%, and ≥99% of all included genomes to cover different levels of conservation. The median centroid counts were computed over all permutations. The curve of the total number of centroids was fitted using nonlinear least-squares estimates (R method “nls”) of the power law function $n = a \cdot N^{\gamma}$ (where n is the total number of centroids, N is the number of included genomes, and a and γ are constants) to the median counts.

Two-dimensional embedding of pan-genome centroids

BusyBee Web [58] was used to represent the pan-genome centroids in two dimensions (2D). In brief, pentanucleotide frequencies were computed and transformed into 2D using Barnes–Hut stochastic neighbor embedding [59]. Due to the use of centroids rather than contigs or long reads, the border point threshold and cluster point threshold were set to 500. Individual pan-genomes were mixed *in silico*, centroids with a frequency ≥90% were used as input to BusyBee Web, and the 2D coordinates were downloaded. Here, in addition to the sample frequency overlay, centroids were colored according to the respective species of the source pan-genome of the centroid.

Resistance association analysis

Association between resistance profiles and centroid presence

All isolates that were used to construct the pan-genomes and had resistance profiles available were considered. Binary centroid presence/absence matrices were used as features. A species–drug combination was not analyzed if > 90% of the isolates were resistant or non-resistant. The predictors were first filtered to remove (nearly) constant and correlated features and features with many missing values. All predictors with > 95% missing values or with > 95% of the entries having the same value (missing values ignored) were removed. Correlated features were removed by computing pairwise feature correlations (fastCor from R-package HiClimR, v. 1.2.3), clustering them using hierarchical clustering (distance = 1 – cor^2, average linkage), cutting the resulting tree at height 0.0975 (1–0.95^2), and keeping only medoids (minimal average distance to other cluster members) within each obtained cluster. All features were scored using EIGENSTRAT [60] (v. 6.0.1) to correct for possible population structures. First, principal component analysis (PCA) was run to compute the top 50 principal components using only retained features. Then, the number of components (k) used for the subsequent computation was chosen such that the estimated genomic inflation factor (lambda) was < 1.1 for the smallest possible k. If none of the computed lambda values was < 1.1, then k with the smallest lambda value was chosen. The value of k was successively increased from k = 1 to k = 50 by an increment of 2. With the chosen value of k, test statistics were generated for all features and P values were computed using the Chi-squared distribution with one degree of freedom. Finally, FDR adjustment was applied.

Number of Resfams covered by the significant resistance association results

For each centroid with a significant resistance association result (adjusted $P < 1E-5$), all hits from the centroid cluster members to the Resfams core database were retrieved. Subsequently, for each Resfam, the number of unique centroids including ≥1 cluster member with a hit to the corresponding Resfam was counted.

Application example

The assembly of the complete *K. pneumoniae* genome published by Kao et al. [39] (NCBI assembly No. ASM195283v1, RefSeq assembly accession No. GCF_001952835.1) was included in the collection of the finished bacterial genomes downloaded from the NCBI RefSeq database as described above. The genomic FASTA file containing the chromosome and plasmid sequences was uploaded to the GEAR-base web-server for genome analysis using default parameters (https://gear-base.com/gear/pangenome/genomesearch/job=b568c458-f68a-4aa1-b78b-dad72dddfd5a/). The FASTA file containing the nucleotide sequences of all CDSs was uploaded for gene-based analysis with only Resfams search and Sourmash search in centroids enabled and using default parameters (https://gear-base.com/gear/pangenome/genesearch/job=0e42e149-a70d-4796-b40a-7f7168dc5077/). The nucleotide sequences of eight resistance genes reported previously [39], including WP_004176269.1, WP_076027158.1, WP_004146118.1, WP_000018329.1, WP_032491824.1, WP_000557454.1, WP_000976514.1, and WP_000027057.1, were saved in a separate FASTA file, which was uploaded for gene-based analysis with all options enabled and default parameters (https://gear-base.com/gear/pangenome/genesearch/job=d8792c0e-bbe7-4936-a7b7-c2846b727afe/).

Availability

GEAR-base is freely available for academic research use after the user has registered and accepted the terms of use available at https://gear-base.com. Because of the sheer size and further legal and ethical constrains, we cannot make all data fully accessible for batch download. If users are interested in getting access to the raw sequencing data, a special request in this respect is required. For this, we provide a respective request details on the GEAR-base homepage. The sequences of pangenome centroids can be downloaded directly from the GEAR-base homepage. Custom scripts used for processing, analyzing and plotting the data can be found at https://github.com/VGalata/gear_base_scripts/.

Authors' contributions

VG performed the computational analysis, implemented the database, and drafted the manuscript together with CCL. CCL and CB also contributed to the data analysis. GH-S and AF performed the next-generation sequencing of the isolates. MH and LvM provided the *S. aureus* isolate collection. AEP, SS, CS, and AP provided the Gram-negative isolate collection. EM, RM, and AK reviewed the manuscript and provided comments. All authors read and approved the final manuscript.

Competing interests

CS and AEP were employees of Siemens Healthcare during the period of the study. SS is an employee of Siemens Healthcare. AEP and AP are Managing Directors of Ares Genetics GmbH, a wholly owned subsidiary of Curetis GmbH. Ares Genetics GmbH is the sole owner of any and all rights to the data presented in the manuscript and in the web resource at https://gear-base.com. Those who are interested in commercial applications or collaboration are invited to contact Ares Genetics at contact@ares-genetics.com.

Acknowledgments

Research for this study was supported by Siemens Healthcare, the Curetis Group, and in parts by the Best Ageing Program from the European Union (Grant No. 306031) as well as the Austrian Research Promotion Agency (Grant Nos 866389 and 863729). We would like to thank Siemens Healthcare and the Curetis Group for their support and for the datasets provided. We are grateful to Laura Smoot, Andrea L. Mrotz, Khoa D. Nguyen, Michael A. Andora, Jose Enrique Fernandez, Nicholas E. Terzakis, Paula Swiatkowski, Usha Vajapey, and Stacie Ho for technical support. We also would like to thank Andy Ying and Gabriel Rensen for their support.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2018.11.002.

References

[1] Bax RP. Antibiotic resistance: a view from the pharmaceutical industry. Clin Infect Dis 1997;24:S151–3.

[2] Rhomberg PR, Jones RN. Summary trends for the meropenem yearly susceptibility test information collection program: a 10-year experience in the United States (1999–2008). Diagn Microbiol Infect Dis 2009;65:414–26.

[3] Center for Disease Dynamics, Economics & Policy. The State of the World's Antibiotics, 2015. [Internet]. Washington DC: Center for Disease Dynamics, Economics & Policy; 2015, http://cddep.org/publications/state_worlds_antibiotics_2015.

[4] World Health Organization. Antimicrobial resistance: global report on surveillance. [Internet]. Geneva, Switzerland: World Health Organization; 2014, http://apps.who.int//iris/handle/10665/112642.

[5] Zhanel GG, DeCorby M, Laing N, Weshnoweski B, Vashisht R, Tailor F, et al. Antimicrobial-resistant pathogens in intensive care units in Canada: results of the Canadian National Intensive Care Unit (CAN-ICU) study, 2005–2006. Antimicrob Agents Chemother 2008;52:1430–7.

[6] Zhanel GG, DeCorby M, Adam H, Mulvey MR, McCracken M, Lagacé-Wiens P, et al. Prevalence of antimicrobial-resistant pathogens in Canadian hospitals: results of the Canadian Ward Surveillance Study (CANWARD 2008). Antimicrob Agents Chemother 2010;54:4684–93.

[7] Karlowsky JA, Lagacé-Wiens PRS, Simner PJ, DeCorby MR, Adam HJ, Walkty A, et al. Antimicrobial resistance in urinary tract pathogens in Canada from 2007 to 2009: CANWARD surveillance study. Antimicrob Agents Chemother 2011;55:3169–75.

[8] European Centre for Disease Prevention and Control. Antimicrobial resistance surveillance in Europe 2015. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net). [Internet]. Stockholm, Sweden: European Centre for Disease Prevention and Control; 2017, https://ecdc.europa.eu/en/publications-data/antimicrobial-resistance-surveillance-europe-2015.

[9] Mendes RE, Castanheira M, Woosley LN, Stone GG, Bradford PA, Flamm RK. Molecular β-lactamase characterization of aerobic Gram-negative pathogens recovered from patients enrolled in the ceftazidime-avibactam phase 3 trials for complicated intra-abdominal infections: Efficacies analyzed against susceptible and resistant subset. Antimicrob Agents Chemother 2017;AAC.02447–16.

[10] Sader HS, Castanheira M, Huband M, Jones RN, Flamm RK. WCK 5222 (cefepime-zidebactam) antimicrobial activity against clinical isolates of Gram-negative bacteria collected worldwide in 2015. Antimicrob Agents Chemother 2017;61:AAC.00072–17.

[11] Castanheira M, Mendes RE, Jones RN, Sader HS. Changes in the frequencies of β-lactamase genes among Enterobacteriaceae isolates in U.S. hospitals, 2012 to 2014: Activity of ceftazidime-avibactam tested against β-lactamase-producing isolates. Antimicrob Agents Chemother 2016;60:4770–7.

[12] Sader HS, Farrell DJ, Flamm RK, Jones RN. Antimicrobial susceptibility of Gram-negative organisms isolated from patients hospitalised with pneumonia in US and European hospitals: Results from the SENTRY Antimicrobial Surveillance Program, 2009–2012. Int J Antimicrob Agents 2014;43:328–34.

[13] Center for Disease Control and Prevention. Antibiotic resistance threats in the United States, 2013. [Internet]. Atlanta, GA: Center for Disease Control and Prevention; 2013, https://www.cdc.gov/drugresistance/pdf/ar-threats-2013-508.pdf.

[14] Köser CU, Ellington MJ, Peacock SJ. Whole-genome sequencing to control antimicrobial resistance. Trends Genet 2014;30:401–7.

[15] Cockerill III FR. Genetic methods for assessing antimicrobial resistance. Antimicrob Agents Chemother 1999; 43:199–212.

[16] Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev 2005;15:589–94.

[17] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015:btv421.

[18] Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter F-J, Zakrzewski M, et al. EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics 2009;10:154.

[19] Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. Nucleic Acids Res 2018;46:e5.

[20] Trost E, Blom J, de Castro Soares S, Huang I-H, Al-Dilaimi A, Schroder J, et al. Pangenomic study of *Corynebacterium diphtheriae* that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. J Bacteriol 2012;194:3199–215.

[21] Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, et al. The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar *ovis* and *equi* strains. PLoS One 2013;8:e53818.

[22] Kant R, Rintahaka J, Yu X, Sigvart-Mattila P, Paulin L, Mecklin J-P, et al. A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in *Lactobacillus rhamnosus*. PLoS One 2014;9:e102762.

[23] De Maayer P, Chan W, Rubagotti E, Venter SN, Toth IK, Birch PRJ, et al. Analysis of the *Pantoea ananatis* pan-genome reveals factors underlying its ability to colonize and interact with plant, insect and vertebrate hosts. BMC Genomics 2014;15:404.

[24] Huang K, Brady A, Mahurkar A, White O, Gevers D, Huttenhower C, et al. MetaRef: a pan-genomic database for comparative and community microbial genomics. Nucleic Acids Res 2014;42: D617–24.

[25] Aanensen DM, Feil EJ, Holden MTG, Dordel J, Yeats CA, Fedosejev A, et al. Whole-genome sequencing for routine pathogen surveillance in public health: a population snapshot of invasive *Staphylococcus aureus* in Europe. MBio 2016;7: e00444–516.

[26] Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, et al. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. J Antimicrob Chemother 2013;68:2234–44.

[27] Strauß L, Ruffing U, Abdulla S, Alabi A, Akulenko R, Garrine M, et al. Detecting *Staphylococcus aureus* virulence and resistance genes: a comparison of whole-genome sequencing and DNA microarray technology. J Clin Microbiol 2016;54:1008–16.

[28] Phaku P, Lebughe M, Strauß L, Peters G, Herrmann M, Mumba D, et al. Unveiling the molecular basis of antimicrobial resistance in *Staphylococcus aureus* from the Democratic Republic of the Congo using whole genome sequencing. Clin Microbiol Infect 2016;22:644.e1–5.

[29] Xavier BB, Das AJ, Cochrane G, De Ganck S, Kumar-Singh S, Aarestrup FM, et al. Consolidating and exploring antibiotic resistance gene data resources. J Clin Microbiol 2016;54:851–9.

[30] Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 2014;42:D581–91.

[31] Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res 2016: gkw1017.

[32] Galata V, Backes C, Laczny CC, Hemmrich-Stanisak G, Li H, Smoot L, et al. Comparing genome versus proteome-based identification of clinical bacterial isolates. Brief Bioinform 2016: bbw122.

[33] Dupont CL, Rusch DB, Yooseph S, Lombardo M-J, Richter RA, Valas R, et al. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J 2012;6:1186–99.

[34] Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J 2015;9:207–16.

[35] Brown CT, Irber L. sourmash: a library for MinHash sketching of DNA. J Open Source Softw 2016;1:27.

[36] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421.

[37] Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 2014;42:D553–9.

[38] Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17:132.

[39] Kao CY, Yan JJ, Lin YC, Zheng PX, Wu JJ. Complete genome sequence of carbapenem-resistant *Klebsiella pneumoniae* strain 1756, isolated from a pus specimen. Genome Announc 2017;5: e00066–117.

[40] Ito R, Mustapha MM, Tomich AD, Callaghan JD, McElheny CL, Mettus RT, et al. Widespread fosfomycin resistance in Gram-negative bacteria attributable to the chromosomal *fosA* gene. MBio 2017;8:e00749–817.

[41] Hennig G, Gehrmann M, Stropp U, Brauch H, Fritz P, Eichelbaum M, et al. Automated extraction of DNA and RNA from a single formalin-fixed paraffin-embedded tissue section for analysis of both single-nucleotide polymorphisms and mRNA expression. Clin Chem 2010;56:1845–53.

[42] Clinical and Laboratory Standards Institute. Performance standards for antimicrobial susceptibility testing; twenty-fourth informational supplement. [Internet]. Wayne, PA: Clinical and Laboratory Standards Institute; 2014, https://www.researchgate.net/publication/307877984_Performance_standards_for_antimicrobial_susceptibility_testing_twenty-fourth_informational_supplement.

[43] European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters; Version 4.0. 2014. [Internet]. Basel, Switzerland: European Committee on Antimicrobial Susceptibility Testing; 2014, http://www.eucast.org/clinical_breakpoints/.

[44] Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;15:R46.

[45] Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R. F1000Research 2013;2:191.

[46] Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2009;37:D5–15.

[47] Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a web browser. BMC Bioinformatics 2011;12:385.

[48] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.

[49] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77.

[50] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9.

[51] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013;29:1072–5.

[52] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.

[53] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.

[54] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2.

[55] Brenner DJ, Fanning GR, Steigerwalt AG, Orskov I, Orskov F. Polynucleotide sequence relatedness among three groups of pathogenic *Escherichia coli* strains. Infect Immun 1972;6:308–15.

[56] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7.

[57] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3.

[58] Laczny CC, Kiefer C, Galata V, Fehlmann T, Backes C, Keller A. BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. Nucleic Acids Res 2017;45: W171–9.

[59] van der Maaten L. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res 2014;15:3221–45.

[60] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–9.

3.3 *BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation*

Published online 2 May 2017 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* ***W171–W179***
doi: 10.1093/nar/gkx348

BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation

Cedric C. Laczny[*,†], Christina Kiefer[†], Valentina Galata, Tobias Fehlmann, Christina Backes and Andreas Keller

Chair for Clinical Bioinformatics, Saarland University, Campus Building E2.1, 66123 Saarbrücken, Germany

Received January 31, 2017; Revised April 11, 2017; Editorial Decision April 18, 2017; Accepted April 21, 2017

ABSTRACT

Metagenomics-based studies of mixed microbial communities are impacting biotechnology, life sciences and medicine. Computational binning of metagenomic data is a powerful approach for the culture-independent recovery of population-resolved genomic sequences, i.e. from individual or closely related, constituent microorganisms. Existing binning solutions often require *a priori* characterized reference genomes and/or dedicated compute resources. Extending currently available reference-independent binning tools, we developed the BusyBee Web server for the automated deconvolution of metagenomic data into population-level genomic bins using assembled contigs (Illumina) or long reads (Pacific Biosciences, Oxford Nanopore Technologies). A reversible compression step as well as bootstrapped supervised binning enable quick turnaround times. The binning results are represented in interactive 2D scatterplots. Moreover, bin quality estimates, taxonomic annotations and annotations of antibiotic resistance genes are computed and visualized. Ground truth-based benchmarks of BusyBee Web demonstrate comparably high performance to state-of-the-art binning solutions for assembled contigs and markedly improved performance for long reads (median F1 scores: 70.02–95.21%). Furthermore, the applicability to real-world metagenomic datasets is shown. In conclusion, our reference-independent approach automatically bins assembled contigs or long reads, exhibits high sensitivity and precision, enables intuitive inspection of the results, and only requires FASTA-formatted input. The web-based application is freely accessible at: https://ccb-microbe.cs.uni-saarland.de/busybee.

INTRODUCTION

Metagenomic sequencing, i.e. whole genome sequencing of DNA indiscriminately extracted from mixed microbial communities, was successfully used to study the taxonomic composition as well as the functional potential of environmental microbiomes (1–4). The independence of prior isolate culturing steps is often considered an advantage as this independence allows reduction in costs and time, as well as the potential to characterize microorganisms that, thus far, have resisted culturing attempts under artificial laboratory conditions (5,6). While metagenomic sequencing has been mostly used for basic research, its potential in clinical settings has been demonstrated recently (7,8). Moreover, third generation-sequencing technologies, e.g. from Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT), are emerging and enable the long read-based study of mixed microbial communities (9–11).

The recovery of genomic sequences resolved at the level of individual organisms (or populations of closely related organisms) from metagenomic sequencing data using computational solutions is termed ‘binning’. The current body of binning approaches can be roughly subdivided into (i) reference-dependent approaches and (ii) reference-independent approaches. Reference-dependent binning approaches are typically characterized by very low runtimes as well as high degrees of sensitivity and precision (12–16). However, these approaches, by design, perform best for sequences derived from organisms that are part of or are closely related to the references present in a database, and are challenged by genomic sequences derived from hitherto uncharacterized microorganisms. In contrast, reference-independent binning approaches do not rely on prior knowledge as they infer sequence cluster structures from the input data only (17–20) and are mostly based on sequence composition, with approaches relying on abundance co-variation across multiple samples emerging recently (21–24). Due to their reference independence, these approaches are of particular use for the analysis of environments with limited representations in the current reference genome databases, frequently allowing resolution

[*]To whom correspondence should be addressed. Tel: +49 681 302 68610; Fax: +49 6841 1626185; Email: cedric.laczny@ccb.uni-saarland.de
[†] These authors contributed equally to the paper as first authors.

of 'unclassified' sequences. However, reference-independent binning often requires substantial amounts of CPU hours, sequence lengths above a certain threshold, e.g. 1000 bp, and/or multiple, ideally independent, samples. While various binning web servers exist, these are mostly based on reference-dependent approaches (15,25–27), or require upfront computations which results in the need for dedicated computing resources and/or user training (28,29).

Here, we extend the currently available reference-independent binning tools by presenting the BusyBee Web server, a web application implementing bootstrapped supervised binning (BSB) of metagenomic sequencing datasets. Our binning approach combines unsupervised and supervised machine learning approaches by 'bootstrapping' the training data from the input rather than relying on reference databases. BusyBee Web only requires a single FASTA-formatted file as input and performs automated deconvolution of the sequences into population-resolved bins. During BSB, clusters are defined *de novo* on a subset of the sequences using an unsupervised approach (30–32). This step is followed by the training of a random forest-based classifier using the cluster labels as the response/dependent variables (supervised part). To further accelerate the binning, an optional 'compression' step is implemented in which data points are randomly sampled serving as representatives for their nearest neighbors (associates) during the unsupervised part (compression). The representatives as well as their associates are subsequently used during the supervised part in combination with the respective representatives' *de novo* cluster labels (decompression). Thus, the training set size is increased compared to only using the randomly sampled, representative data points. Ultimately, every sequence ($\geq$500 bp, by default) is assigned a label using the bootstrap-trained classifier, thereby defining the final set of bins. For inspection of the clustering/binning results, a 2D scatterplot of the data-inherent as well as the inferred structures is presented to the user. To complement this, estimates of bin quality, i.e. degrees of completeness, contamination and strain heterogeneity, are computed and visualized. Moreover, sequences are taxonomically annotated using Kraken and functional annotation of antibiotic resistance genes is performed. Because all of the binning and annotation steps are automatically executed by the web server transparently to the user, no dedicated computing resources or special user training is required. Furthermore, custom per-sequence annotations can be uploaded by the user, e.g. to highlight specific sequences of interest, and BusyBee Web offers the option to download the generated results should specialized downstream analyses be required, e.g. population-resolved annotation of KEGG pathways. Ground truth-based benchmarks comparing our BSB approach to state-of-the-art binning approaches are provided for assembled contigs (Illumina) and long reads (ONT). Moreover, the applicability of our web server for the analysis of real-world metagenomic datasets (Illumina or PacBio) is demonstrated. The BusyBee Web server is available free-to-use at https://ccb-microbe.cs.uni-saarland.de/busybee.

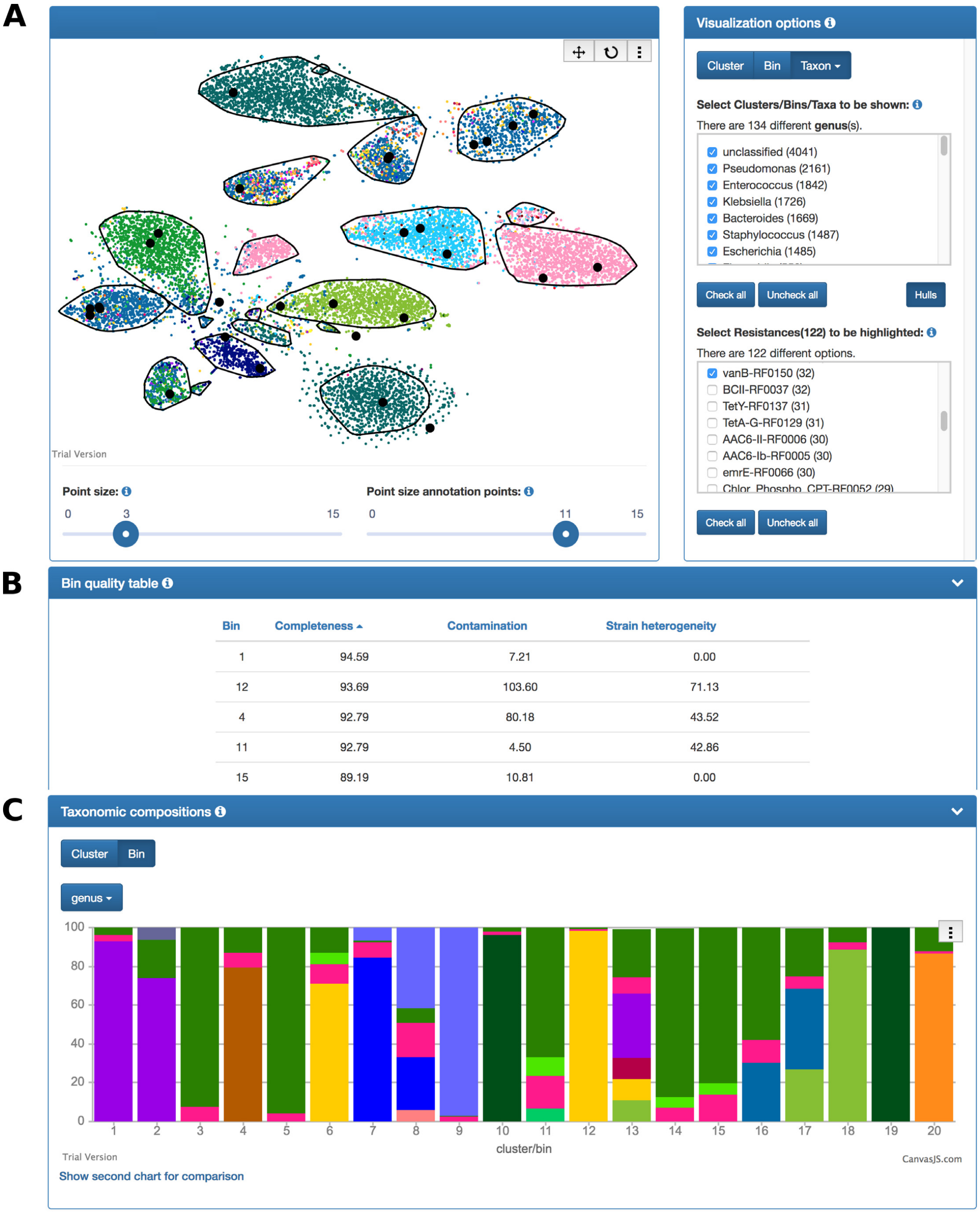
IMPLEMENTATION

Workflow

When a new job is initiated, the user has to provide a FASTA-formatted file of nucleotide sequences, e.g. assembled contigs or long sequencing reads, as the only mandatory input. By default, population-level genomic bins are automatically defined by BSB of the input sequences followed by bin quality assessment. Moreover, BusyBee Web can optionally compute taxonomic annotations and annotations of antibiotic resistance genes. Custom, per-sequence annotations can also be provided by the user, e.g. to highlight specific sequences of interest. Importantly, as BSB is a reference-independent approach, population-level resolution is achieved even in the absence of taxonomically annotated reference genomes, e.g. for environments with limited representations in current reference databases. Robust default values for all BSB parameters are pre-set but can be adjusted by the user. Upon completion of all computational steps, the user can explore the results through interactive visualizations directly in the browser (HTML, JavaScript), e.g. to identify bins that are enriched for specific antibiotic resistance genes or bins that represent candidate hitherto uncharacterized microorganisms (Figure 1). Individual results can be shared using the unique job ID or the URL of the results page. Moreover, a zipped archive of the results can be downloaded for downstream processing. This archive includes the binning results (in particular, per-bin FASTA files), results from the bin quality assessment, as well as results from the optional taxonomic and functional annotation steps.

Bootstrapped supervised binning

BSB is reference-independent and combines unsupervised as well as supervised machine learning approaches using genomic signatures in the form of oligonucleotide frequencies as the feature set (Supplementary Materials and Supplementary Figure S1). Supervised binning approaches are often equated with reference-dependent approaches, in particular, using reference genomes derived from microbial isolates for a priori training. However, more generally, a supervised machine learning approach uses training data to generate a model which is subsequently used for the classification of test data. The training data can be inferred from the input data as it is effectively done in our approach by first using an unsupervised machine learning approach (Supplementary Figure S1A). Accordingly, BSB can be seen as an extension of a classifier trained on a specific set of references, albeit bootstrapping the training data from the input data (19), rather than relying on previously characterized reference sequences (13,33). In brief, sequences are size-selected and separated into border points, cluster points and remaining points according to the user-specified parameters (border points sequence length threshold, t_b and the cluster points sequence length threshold, t_c). Each sequence is then represented by its genomic signature, using pentanucleotide frequencies (default). Optionally, the border points and the cluster points are 'compressed' which will reduce the runtimes of both, the 2D embedding and the automated clustering. Following this, the 2D embedding is computed



Point size: ℹ 0 3 15

Point size annotation points: ℹ 0 11 15

Visualization options ℹ

Cluster | Bin | Taxon ▾

Select Clusters/Bins/Taxa to be shown: ℹ

There are 134 different **genus**(s).

☑ unclassified (4041)
☑ Pseudomonas (2161)
☑ Enterococcus (1842)
☑ Klebsiella (1726)
☑ Bacteroides (1669)
☑ Staphylococcus (1487)
☑ Escherichia (1485)

Check all | Uncheck all | Hulls

Select Resistances(122) to be highlighted: ℹ

There are 122 different options.

☑ vanB-RF0150 (32)
☐ BCII-RF0037 (32)
☐ TetY-RF0137 (31)
☐ TetA-G-RF0129 (31)
☐ AAC6-II-RF0006 (30)
☐ AAC6-Ib-RF0005 (30)
☐ emrE-RF0066 (30)
☐ Chlor_Phospho_CPT-RF0052 (29)

Check all | Uncheck all

Bin quality table ℹ

Bin	Completeness ▴	Contamination	Strain heterogeneity
1	94.59	7.21	0.00
12	93.69	103.60	71.13
4	92.79	80.18	43.52
11	92.79	4.50	42.86
15	89.19	10.81	0.00

Taxonomic compositions ℹ

Cluster | Bin

genus ▾

Show second chart for comparison

Figure 1. Overview of individual components of the BusyBee Web results page. (**A**) Input sequences are represented as individual points (according to the thresholds t_b and t_c) in the 2D scatterplot. Convex hulls (black polygons) delineate the predicted clusters. If the optional taxonomic and functional annotations were enabled, taxon and antibiotic resistance-related information is shown to the right of the scatterplot. Individual clusters, bins or taxa can be shown or hidden and sequences encoding for specific antibiotic resistance genes can be highlighted using points of larger size and dark color, here, for the *vanB* gene. A left-click on a point reveals detailed information about the respective sequence, e.g. the taxonomic lineage or encoded antibiotic resistance genes. The user can pan and zoom the plot using the mouse, e.g. to focus on a region of interest, and point sizes are easily adjusted using sliders below the 2D scatterplot. (**B**) Bin quality estimates (completeness, contamination, strain heterogeneity) are provided as a sortable table, here, sorted by decreasing completeness. An excerpt representing the five most complete bins is shown. (**C**) The optional taxonomic compositions of the clusters/bins are shown as stacked bar charts. The taxonomic rank, e.g. genus, can be selected and a second chart can be shown to compare the compositions of the individual clusters/bins at different ranks, e.g. genus versus family.

(20,32). Subsequently, automated clustering (30,31) is performed on the cluster points only, while the border points are supposed to help push individual clusters further apart and, thus, improve the automated segregation into distinct sequence clusters (Supplementary Figure S1B). The clustering information is used to train a random forest-based classifier, which predicts cluster assignments for the input sequences (≥500 bp; default), thereby defining the final set of bins (Supplementary Figure S1). In this context, it is important to highlight the difference between a 'cluster' and a 'bin'. While both represent sequence sets, a 'cluster' is an intermediate sequence set and a 'bin' is a final sequence set. Consequently, a cluster may represent only a limited fraction of a population-level genome, while a bin tries to maximize the recovery of genomic information derived from the respective population.

While generally robust default parameters are provided in BusyBee Web, the user might need to specify custom settings based on the characteristics of the input data. For example, given highly fragmented assemblies or datasets with narrow sequence length distributions, the border points sequence length threshold, t_b and the cluster points sequence length threshold, t_c, may be set to equal values, e.g. decreasing t_c to the value of t_b. The 'minPts' parameter value may be decreased to allow the identification of small-sized clusters. In this context, if the degree of compression is set too high and the 'minPts' parameter is not decreased accordingly, clusters might be missed. This typically becomes apparent as distinct groups of points in the 2D visualization lacking a convex hull, i.e. not delimited by a polygon. Moreover, increasing the minimum sequence length can avoid the annotation of short sequences and thus decrease the fraction of incomplete genes. Detailed parameter descriptions are provided in the Supplementary Materials and as online tooltips.

Annotations

Taxonomic annotation. Kraken (v0.10.5-beta) in combination with the Minikraken database, i.e. a reduced-size database constructed from complete bacterial, archaeal and viral genomes in RefSeq as of 8 December 2014 (https://ccb.jhu.edu/software/kraken/dl/minikraken.tgz), is used to compute taxonomic annotations for the input sequences (14). The reduced-size database was chosen due to its low memory requirements. However, the integration of a larger database is possible in the future to increase the sensitivity of the taxonomic annotations.

Annotation of antibiotic resistance genes. Prokka (v1.11) with the '—fast' option is used for gene (CDS) calling (34,35) on all input sequences. The translated CDS sequences are then searched against the ResFams collection of antibiotic resistance genes (36) using hmmsearch from HMMER (v3.1b2; http://hmmer.janelia.org/).

Custom annotations. Custom annotations can be uploaded to highlight individual sequences or sequence sets. The former can, for example, be used for sequences encoding genes with a particular function and the latter for sequences annotated with a custom reference genome database or characterized according to their genomic or transcriptomic fold-coverage, or ratio of both (high/medium/low) (37). To enable this option, a tab-separated text file containing the sequence ID in the first column and the respective annotation in the second column should be provided by the user.

Bin quality assessment. CheckM (v1.0.7) is used to evaluate the quality (degrees of completeness, contamination and strain heterogeneity) of the individual bins using a custom set of marker genes ('essential genes') (38–40). The default memory requirement of CheckM (≥16 GB RAM) is prohibitive for use in a web application serving multiple users concurrently. Hence, the use of a custom set of marker genes which reduces the memory requirements of CheckM considerably by bypassing the reference genome-tree placement. While the currently implemented custom set is bacteria-specific, extended sets can be integrated into BusyBee Web in the future to represent microorganisms from other domains, e.g. archaea.

Representation of the results

BusyBee Web provides interactive visualizations of the results (Figures 1 and 2). The automated clustering/binning results are represented as a 2D scatterplot, with individual points colored according to their assignment (cluster, bin or noise; Supplementary Figure S1) and each point representing an input sequence with length $\geq t_b$. Convex hulls are additionally plotted to help in delineating the individual clusters. Suboptimal automatically defined clusters can thus be identified visually, e.g. distinct clusters which have been artificially joined. Clicking on individual points provides detailed information on the point's optional annotations, i.e. the predicted taxonomy and antibiotic resistance genes encoded by the respective sequence. Moreover, the user can change the size of the points as well as pan and zoom the plot. Individual clusters, bins or taxonomic groups (e.g. at the genus-level or at the species-level) can be selected. Unselected points are plotted with reduced opacity. Similarly, groups of sequences encoding specific antibiotic resistance genes or sharing individual, user-provided annotation can be shown or hidden. The number of contigs per cluster and per bin is shown as a bar chart. This allows the user to see how many sequences represented a cluster during the training phase and how many sequences were assigned to a bin by the trained classifier. Furthermore, bin quality estimates (completeness, contamination, strain heterogeneity) are displayed as a bar chart and as a sortable table. The taxonomic compositions per cluster and per bin are shown as stacked percent bar charts and a second chart of taxonomic compositions can be opened, thereby allowing the comparison of cluster/bin taxonomic compositions at different taxonomic ranks, e.g. at the family-level and the genus-level. A zipped archive of the results, including per-bin FASTA-formatted files and per-sequence taxonomic annotations among others, can be downloaded.

Metagenomic datasets to evaluate BusyBee Web

Two metagenomic datasets of short read-assembled contigs (Shakya2013 (41), Gregor2016 (13); Illumina) as well as one

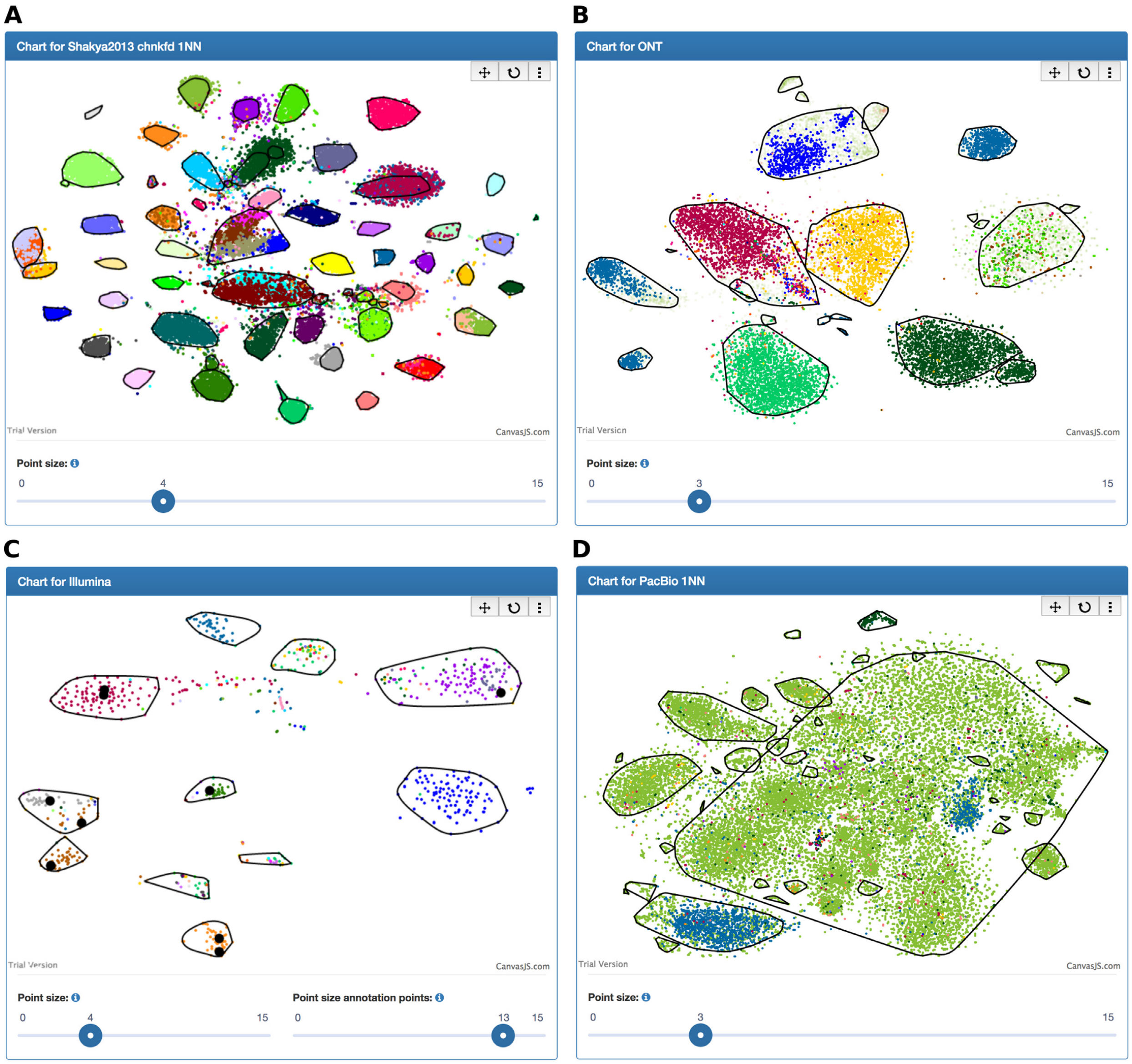


Figure 2. Screenshots of the interactive scatterplots for (**A**) ground truth-based Illumina (Shakya2013), (**B**) ground truth-based ONT, (**C**) small-scale Illumina and (**D**) PacBio metagenomic data. (A) A compression of 1 ('1NN') as well as sequence chunks (3 kbp chunk-length) derived from the full-length contigs were used. (B) Only sequences with species-level taxonomic assignments are shown. (C) Sequences encoding for class A CTX-M beta-lactamases (CTXM-RF0059) are highlighted. (D) A compression of 1 ('1NN') was used. The convex hulls (black polygons) delineate the individual sequence clusters. Descriptions at the top of each plot represent job names; if none is specified, a unique job ID is shown. Colors are based on species-level taxonomic assignments.

raw, long read sequencing-based dataset (ONT) representing microbial communities of known composition (42–48) (Table 1, Supplementary Materials and Supplementary Table S1), i.e. representing ground truth data, were used to quantitatively assess the performance of our BSB approach and to compare it against two state-of-the-art binning approaches, MaxBin2 and MetaBAT (23,24).

Three additional metagenomic datasets were used to demonstrate the versatility of the BusyBee Web server: two Illumina-based datasets (small-scale (49), large-scale (39)) and a PacBio-based dataset (10) were used as originally provided (Table 1). The PacBio dataset consisted of Circular Consensus Sequences (CCS) which provide increased sequence quality by repeatedly sequencing the same molecule (50), thereby correcting for sequence errors. However, no additional error correction nor assembly were performed on the CCS reads.

Table 1. BusyBee Web runtimes reported in minutes for the herein studied ground truth and real-world datasets

		# sequences	Total length [bp]	Binning runtime [min]	Total runtime [min]
Ground truth	Shakya2013	24 974	179 063 212	8	30
	Gregor2016	14 393	142 556 476	6	23
	ONT	21 000	97 715 136	11	20
Real-world	Small-scale Illumina	859	50 964 782	1	6
	Large-scale Illumina‡	133 149	399 132 179	28	75
	PacBio†	71 029	93 937 106	18	27

Runtimes were determined manually based on the progress interface in the browser and were rounded to the next full minute. The minimum sequence length threshold was 1 kbp for the large-scale Illumina dataset and 500 bp for the other datasets.
†Compression of 1 was used.
‡Compression of 2 was used.

RESULTS AND DISCUSSION

To cover the heterogeneity of currently available sequencing technologies, we applied BusyBee Web to Illumina-, PacBio- and ONT-based sequencing data. Moreover, we compared the binning performance of BusyBee Web against MaxBin2 and MetaBAT on three ground truth datasets.

Ground truth-based evaluation of BSB

We used two Illumina-based (Shakya2013, Gregor2016) and one ONT-based metagenomic dataset of defined composition to evaluate our BSB approach (Table 1). The numbers of bins inferred by BSB were 45/38 (Shakya2013/Gregor2016), with 58/45 expected species (Supplementary Notes and Supplementary Tables S2–5). Normalization of the cluster density by using sequence chunks (3 kbp chunk-length) derived from the full-length contigs (49,51) resulted in 60/50 bins. Moreover, the sensitivity, precision and F1 values were substantially increased (Supplementary Tables S3 and 5). For example, the median precision value was almost 20% higher using sequence chunks (91.49%; Figure 2A) instead of the full-length contigs (71.99%; Supplementary Figure S2), and the median F1 score increased to 90.09 from 70.02% for the Shakya2013 dataset.

For the ONT-based ground truth data ($t_b = t_c = 500$ bp), our approach reported 23 bins, with the large bins representing the six constituent bacterial organisms (Figure 2B). The influenza A virus-derived sequences formed at least three major bins. The bin quality assessment yielded no representative results which may be due to the increased error-rate of the raw, nanopore sequencing-based reads (52–54) and the use of read subsamples for this dataset (Supplementary Table S6). About 31.91% (6701/21 000) of the sequences remained unclassified at the phylum-level, which is likely due to their increased error-rate. Nevertheless, BusyBee Web created representative bins for all the included isolates resulting in mean/median F1 scores of 89.00/92.66% (Supplementary Table S7). Processing only the influenza A virus-derived (subsampled) sequences revealed eight bins (Supplementary Figure S3). While an in-depth study of the individual bins was beyond the scope of the current work, this serves as an example of using BusyBee Web to inspect microbial isolate-derived genomic sequences or bins generated by a complementary binning tool for the presence of multiple sequence clusters, e.g. due to multiple chromosomes or possible contaminations.

Benchmarking against existing binning tools using ground truth data

We compared the results of our BSB approach to two state-of-the-art approaches, MaxBin2 and MetaBAT. These tools were selected as they both support single sample-based binning. As described above, BSB identified 45/38 (Shakya2013/Gregor2016) bins for the Illumina-based ground truth data. In comparison, MetaBAT produced 63/41 bins and MaxBin2 produced 58 bins for the Shakya2013 data but was omitted for the Gregor2016 data due to missing coverage information. While MetaBAT typically had high precision values for both Illumina-based ground truth datasets, the sensitivity was often low (Supplementary Tables S3 and 5). MaxBin2 had higher mean/median sensitivity compared to MetaBAT on the Shakya2013 data, yet had low mean/median precision (59.76/57.09%). Using our BSB approach, the highest median F1 scores were reached with 90.09 and 95.21% for the Shakya2013 and Gregor2016 chunked datasets, respectively (Supplementary Notes).

For the ONT data, MetaBAT and MaxBin2 returned 18 and 2 bins, resulting in mean/median F1 scores of 58.35/56.92% and 40.26/34.56%, respectively (Supplementary Table S7). MaxBin2 and MetaBAT use an empirically determined probability distribution for the tetranucleotide frequency distances (23,24). This distribution is learned *a priori* on high quality reference genomes. The increased sequence error rate of third generation-sequencing data is likely to negatively impact the distance calculations, i.e. two sequences might have larger tetranucleotide frequency distances despite being derived from the same genome. Consequently, this is likely to have negatively affected the binning performance in MaxBin2 and MetaBAT. Moreover, coverage values are a mandatory input to MaxBin2, yet were unavailable for the unassembled, long read ONT data. Hence, surrogate, unit coverage values were used for MaxBin2 while MetaBAT defaulted to coverage-free binning using tetranucleotide frequencies. While coverage information provides important information for binning and bin refinement (1,21,23,24,51), an initial assembly is required onto which reads can be mapped to compute the fold-coverage of the assembled contigs. However, if the reads are sufficiently long, e.g. >1000 bp, the binning can be performed prior to the assembly, thereby facilitating population-level assemblies. Accordingly, our BSB approach can be used to pre-partition raw, metagenomic, long reads, thus enabling a 'divide and conquer' approach.

Real-world metagenomes

For the small-scale Illumina-based dataset (Figure 2C and Table 1), a total of 11 bins was identified with 7 near-complete bins (≥90% complete) and 4 partially complete bins (≥50%). A total of 5 of the 11 bins had contamination degrees ≥20% with 2 of the 5 bins showing high degrees of strain heterogeneity (≥80%). This indicates that sequences derived from closely related organisms were grouped together while sequences derived from more distantly related organisms were separated. Class A CTX-M β-lactamases were highlighted to demonstrate the antibiotic resistance gene annotation functionality (Figure 2C). A total of 6 of the 11 bins were found to contain sequences encoding for the respective genes. For the large-scale Illumina-based dataset (Table 1) a compression of 2 was used, resulting in 51 bins (Supplementary Figure S6) of which 11 were ≥90% complete and 33 were ≥50% complete. The average/median degrees of completeness, contamination and strain heterogeneity were found to be 62.36/74.77%, 85.96/11.71% and 22.75/9.81%, respectively.

For the analysis of the PacBio dataset (Table 1), a compression of 1 was used and the border points and cluster points thresholds were set to 500 bp due to the small average read length of 1319 bp. A large bin (bin number 1), including sub-structures that were not resolved by the automatic clustering step, dominated the results visualization (Figure 2D; center of the scatterplot). However, the interactive visualization in BusyBee Web enables the user to easily identify suspect bins, e.g. bins with suboptimal automated deconvolution. A detailed inspection and refinement (55–57) of the suspect bins can be subsequently performed using a user-driven binning approach, such as anvi'o or VizBin (29,32). Overall, the bins were less complete compared to the Illumina-based dataset (3 bins ≥50%). It should be noted that the Illumina-based data was derived from the sequencing of 11 samples (∼2.4 Gbp per sample) (49), while the PacBio-based data consisted of 94 Mbp of CCS reads derived from 8 flow cells (10). About 90.82% (32 255/35 515; after compression) of the sequences could not be classified at the phylum level using Kraken in combination with the Minikraken database. However, our BSB approach assigned 91.16% (64 753/71 029) of the total of sequences to the five largest bins.

The total runtimes for the herein studied datasets were between 6 and 75 min (Table 1) and the BSB step required <30 min for the largest dataset (133 149 sequences; 399 132 179 bp). While the taxonomic annotation step is fast (below 5 min for the large-scale Illumina dataset), a considerable and highly variable proportion of the runtime is spent by the bin quality control. The high variability might be explained by varying amounts of identified single copy marker genes.

CONCLUSION

Metagenomic sequencing has become a widely used approach for the culture-independent study of mixed microbial communities and is often coupled with *in silico* deconvolution of metagenomic sequence fragments into population-resolved genomic bins ('binning') in order to study the constituent micro-organisms at an organismal level. While several binning approaches have been developed, they mostly require previously characterized references, substantial computing resources and/or prior user training. Here, we presented the BusyBee Web server for the automated, reference-independent binning and visualization of metagenomic data in the form of assembled contigs (Illumina) or long reads (PacBio, ONT). The web-based interactive representations, including a 2D embedding of genomic signatures, bin quality assessment using single copy marker genes and optional taxonomic assignments, allow for intuitive inspection of the results. This can help the user to build confidence in the individual bins while simultaneously facilitating the identification of sequence groups requiring special attention. In addition, automatically generated annotations of antibiotic resistance gene-encoding sequences or user-provided, per-sequence annotations are optionally overlaid on the 2D embeddings, e.g. with the former allowing to identify population-level genomes enriched for genes possibly conveying specific antibiotic resistances. The only mandatory input consists of a FASTA-formatted nucleotide sequence file and all computations are performed online and transparently to the user. Hence, no special user training, software installation or dedicated computing resources are required and individual results can easily be shared via the web. Moreover, BusyBee Web was evaluated on ground truth and real-world metagenomic data, with the ground truth-based benchmarks demonstrating comparable performance to state-of-the-art binning approaches for assembled contigs and markedly improved performance for long reads when using our approach. Overall, BusyBee Web facilitates population-level resolved analyses of metagenomic data, thereby being of service for the study of mixed microbial communities derived from various environments and sequencing technologies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Anna Heintz-Buschart and Patrick May for fruitful discussions on clustering and classification. Linda Wampach is thanked for the creation of the BusyBee Web logo. We thank Anne Kaysen, Emilie Muller, Malte Herold, Linda Wampach, Anna Heintz-Buschart, Patrick May, Tim Kehl, Alexander Fauss and Kevin Hoffeld for testing the web server and providing feedback.

FUNDING

Saarland University Internal Funds; European Union (Best Ageing) [306031 to V.G., T.F.]. Funding for open access charge: Saarland University.

Conflict of interest statement. None declared.

REFERENCES

1. Muller,E.E.L., Pinel,N., Laczny,C.C., Hoopmann,M.R., Narayanasamy,S., Lebrun,L.A., Roume,H., Lin,J., May,P.,

Hicks,N.D. *et al.* (2014) Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat. Commun.*, **5**, 5603.

2. Howe,A.C., Jansson,J.K., Malfatti,S.A., Tringe,S.G., Tiedje,J.M. and Brown,C.T. (2014) Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 4904–4909.
3. McGarvey,K.M., Queitsch,K. and Fields,S. (2012) Wide variation in antibiotic resistance proteins identified by functional metagenomic screening of a soil DNA library. *Appl. Environ. Microbiol.*, **78**, 1708–1714.
4. Hernández,E., Bargiela,R., Diez,M.S., Friedrichs,A., Pérez-Cobas,A.E., Gosalbes,M.J., Knecht,H., Martínez-Martínez,M., Seifert,J., Von Bergen,M. *et al.* (2013) Functional consequences of microbial shifts in the human gastrointestinal tract linked to antibiotic treatment and obesity. *Gut Microbes.*, **4**, 306–315.
5. Iverson,V., Morris,R.M., Frazar,C.D., Berthiaume,C.T., Morales,R.L. and Armbrust,E.V. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science*, **335**, 587–590.
6. Vartoukian,S.R., Palmer,R.M. and Wade,W.G. (2010) Strategies for culture of 'unculturable' bacteria. *FEMS Microbiol. Lett.*, **309**, 1–7.
7. Loman,N.J., Constantinidou,C., Christner,M., Rohde,H., Chan,J.Z.-M., Quick,J., Weir,J.C., Quince,C., Smith,G.P., Betley,J.R. *et al.* (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *JAMA*, **309**, 1502–1510.
8. van der Helm,E., Imamovic,L., Hashim Ellabaan,M.M., van Schaik,W., Koza,A. and Sommer,M.O.A. (2017) Rapid resistome mapping using nanopore sequencing. *Nucleic Acids Res.*, doi:10.1093/nar/gkw1328.
9. Tsai,Y.-C., Conlan,S., Deming,C. and NISC Comparative Sequencing ProgramNISC Comparative Sequencing Program, Segre,J.A., Kong,H.H., Korlach,J. and Oh,J. (2016) Resolving the complexity of human skin metagenomes using single-molecule sequencing. *Mbio*, **7**, doi:10.1128/mBio.01948-15.
10. Frank,J.A., Pan,Y., Tooming-Klunderud,A., Eijsink,V.G.H., McHardy,A.C., Nederbragt,A.J. and Pope,P.B. (2016) Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci. Rep.*, **6**, 25373.
11. Brown,B.L., Watson,M., Minot,S.S., Rivera,M.C. and Franklin,R.B. (2017) MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. *Gigascience*, **6**, 1–10.
12. Rosen,G.L., Reichenberger,E.R. and Rosenfeld,A.M. (2011) NBC: the naive Bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, **27**, 127–129.
13. Gregor,I., Dröge,J., Schirmer,M., Quince,C. and McHardy,A.C. (2016) PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ*, **4**, e1603.
14. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
15. Menzel,P., Ng,K.L. and Krogh,A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.*, **7**, 11257.
16. Kim,D., Song,L., Breitwieser,F.P. and Salzberg,S.L. (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.*, **26**, 1721–1729.
17. Teeling,H., Waldmann,J., Lombardot,T., Bauer,M. and Glöckner,F.O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, **5**, 163.
18. Dick,G.J., Andersson,A.F., Baker,B.J., Simmons,S.L., Thomas,B.C., Yelton,A.P. and Banfield,J.F. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol.*, **10**, R85.
19. Strous,M., Kraft,B., Bisdorf,R. and Tegetmeyer,H.E. (2012) The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front. Microbiol.*, **3**, 410.
20. Laczny,C.C., Pinel,N., Vlassis,N. and Wilmes,P. (2014) Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci. Rep.*, **4**, 4516.
21. Alneberg,J., Bjarnason,B.S., de Bruijn,I., Schirmer,M., Quick,J., Ijaz,U.Z., Lahti,L., Loman,N.J., Andersson,A.F. and Quince,C. (2014) Binning metagenomic contigs by coverage and composition. *Nat. Methods*, **11**, 1144–1146.
22. Nielsen,H.B., Almeida,M., Juncker,A.S., Rasmussen,S., Li,J., Sunagawa,S., Plichta,D.R., Gautier,L., Pedersen,A.G., Le Chatelier,E. *et al.* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.*, **32**, 822–828.
23. Wu,Y.W., Simmons,B.A. and Singer,S.W. (2015) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, **32**, 605–607.
24. Kang,D.D., Froula,J., Egan,R. and Wang,Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **3**, e1165.
25. Sharma,V.K., Kumar,N., Prakash,T. and Taylor,T.D. (2012) Fast and accurate taxonomic assignments of metagenomic sequences using metabin. *PLoS One*, **7**, e34030.
26. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
27. Mohammed,M.H., Ghosh,T.S., Singh,N.K. and Mande,S.S. (2011) SPHINX-an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics*, **27**, 22–30.
28. Cantor,M., Nordberg,H., Smirnova,T., Hess,M., Tringe,S. and Dubchak,I. (2015) Elviz—exploration of metagenome assemblies with an interactive visualization tool. *BMC Bioinformatics*, **16**, 130.
29. Eren,A.M., Esen,Ö.C., Quince,C., Vineis,J.H., Morrison,H.G., Sogin,M.L. and Delmont,T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, **3**, e1319.
30. Gisbrecht,A., Hammer,B., Mokbel,B. and Sczyrba,A. (2013) Nonlinear dimensionality reduction for cluster identification in metagenomic samples. In: *17th International Conference on Information Visualisation*. IEEE, London, doi:10.1109/IV.2013.22.
31. Heintz-Buschart,A., May,P., Laczny,C.C., Lebrun,L.A., Bellora,C., Krishna,A., Wampach,L., Schneider,J.G., Hogan,A., de Beaufort,C. *et al.* (2016) Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.*, **2**, 16180.
32. Laczny,C.C., Sternal,T., Plugaru,V., Gawron,P., Atashpendar,A., Margossian,H.H., Coronado,S., der Maaten,L.van., Vlassis,N. and Wilmes,P. (2015) VizBin—an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, **3**, 1.
33. McHardy,A.C., Martín,H.G., Tsirigos,A., Hugenholtz,P. and Rigoutsos,I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, **4**, 63–72.
34. Hyatt,D., Chen,G.-L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
35. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
36. Gibson,M.K., Forsberg,K.J. and Dantas,G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, **9**, 207–216.
37. Narayanasamy,S., Jarosz,Y., Muller,E.E.L., Heintz-Buschart,A., Herold,M., Kaysen,A., Laczny,C.C., Pinel,N., May,P. and Wilmes,P. (2016) IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.*, **17**, 260.
38. Dupont,C.L., Rusch,D.B., Yooseph,S., Lombardo,M.-J., Richter,R.A., Valas,R., Novotny,M., Yee-Greenbaum,J., Selengut,J.D., Haft,D.H. *et al.* (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.*, **6**, 1186–1199.
39. Albertsen,M., Hugenholtz,P., Skarshewski,A., Nielsen,K.L., Tyson,G.W. and Nielsen,P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, **31**, 533–538.
40. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
41. Shakya,M., Quince,C., Campbell,J.H., Yang,Z.K., Schadt,C.W. and Podar,M. (2013) Comparative metagenomic and rRNA microbial

diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.*, **15**, 1882–1899.

42. Deschamps,S., Mudge,J., Cameron,C., Ramaraj,T., Anand,A., Fengler,K., Hayes,K., Llaca,V., Jones,T.J. and May,G. (2016) Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from Agrobacterium tumefaciens. *Sci. Rep.*, **6**, 28625.
43. Wang,J., Moore,N.E., Deng,Y.M., Eccles,D.A. and Hall,R.J. (2015) MinION nanopore sequencing of an influenza genome. *Front. Microbiol.*, **6**, 1–7.
44. Karlsson,E., Lärkeryd,A., Sjödin,A., Forsman,M. and Stenberg,P. (2015) Scaffolding of a bacterial genome using MinION nanopore sequencing. *Sci. Rep.*, **5**, 11996.
45. Loman,N.J., Quick,J. and Simpson,J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, **12**, 733–735.
46. Quick,J., Ashton,P., Calus,S., Chatt,C., Gossain,S., Hawker,J., Nair,S., Neal,K., Nye,K., Peters,T. *et al.* (2015) Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.*, **16**, 114.
47. Judge,K., Hunt,M., Reuter,S., Tracey,A., Quail,M.A., Parkhill,J. and Peacock,S.J. (2016) Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. *Microb. Genomics*, **2**, e000085.
48. Bradley,P., Gordon,N.C., Walker,T.M., Dunn,L., Heys,S., Huang,B., Earle,S., Pankhurst,L.J., Anson,L., de Cesare,M. *et al.* (2015) Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.*, **6**, 10063.
49. Sharon,I., Morowitz,M.J., Thomas,B.C., Costello,E.K., Relman,D.A and Banfield,J.F. (2013) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.*, **23**, 111–120.
50. Travers,K.J., Chin,C.S., Rank,D.R., Eid,J.S. and Turner,S.W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.*, **38**, e159.
51. Laczny,C.C., Muller,E.E.L., Heintz-Buschart,A., Herold,M., Lebrun,L.A., Hogan,A., May,P., de Beaufort,C. and Wilmes,P. (2016) Identification, recovery, and refinement of hitherto undescribed population-level genomes from the human gastrointestinal tract. *Front. Microbiol.*, **7**, 884.
52. Laver,T., Harrison,J., O'Neill,P.A., Moore,K., Farbos,A., Paszkiewicz,K. and Studholme,D.J. (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.*, **3**, 1–8.
53. Kilianski,A., Haas,J.L., Corriveau,E.J., Liem,A.T., Willis,K.L., Kadavy,D.R., Rosenzweig,C.N. and Minot,S.S. (2015) Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *Gigascience*, **4**, 12.
54. Jain,M., Fiddes,I.T., Miga,K.H., Olsen,H.E., Paten,B. and Akeson,M. (2015) Improved data analysis for the MinION nanopore sequencer. *Nat. Methods*, **12**, 351–356.
55. Tai,V., Carpenter,K.J., Weber,P.K., Nalepa,C.A., Perlman,S.J. and Keelinga,P.J. (2016) Genome evolution and nitrogen fixation in bacterial ectosymbionts of a protist inhabiting wood-feeding cockroaches. *Appl. Environ. Microbiol.*, **82**, 4682–4695.
56. Buongiorno,J., Bird,J.T., Krivushin,K., Oshurkova,V., Shcherbakova,V., Rivkina,E.M., Lloyd,K.G. and Vishnivetskaya,T.A. (2016) Draft genome sequence of antarctic methanogen enriched from dry valley permafrost. *Genome Announc.*, **4**, 1–2.
57. Russell,J.A., León-Zayas,R., Wrighton,K. and Biddle,J.F. (2016) Deep subsurface life from north pond: enrichment, isolation, characterization and genomes of heterotrophic bacteria. *Front. Microbiol.*, **7**, 678.

3.4 *Assessing the heterogeneity of in silico plasmid predictions based on whole-genome-sequenced clinical isolates*

This is a pre-copyedited, author-produced version of an article accepted for publication in *Briefings in Bioinformatics* following peer review. The version of record *Briefings in Bioinformatics, Volume 20, Issue 3, May 2019, Pages 857–865* is available online at: https://doi.org/10.1093/bib/bbx162.

Assessing the heterogeneity of *in silico* plasmid predictions based on whole-genome-sequenced clinical isolates

Cedric C. Laczny, Valentina Galata, Achim Plum, Andreas E. Posch and Andreas Keller

Corresponding author: Andreas Keller, Saarland University, Building E2.1, 66123 Saarbrücken, Germany. Tel.: +49 (174) 1684638; Fax: +49 (0) 6841-162-6185. E-mail: andreas.keller@ccb.uni-saarland.de.

Cedric C. Laczny is a Postdoc at the Chair for Clinical Bioinformatics at Saarland University.
Valentina Galata is a PhD student at the Chair of Clinical Bioinformatics at Saarland University.
Achim Plum is a Managing Director of Ares Genetics GmbH and Curetis GmbH.
Andreas E. Posch is a Managing Director of Ares Genetics GmbH.
Andreas Keller is a professor and head of the Chair for Clinical Bioinfo

Abstract

High-throughput next-generation shotgun sequencing of pathogenic bacteria is growing in clinical relevance, especially for chromosomal DNA-based taxonomic identification and for antibiotic resistance prediction. Genetic exchange is facilitated for extrachromosomal DNA, e.g. plasmid-borne antibiotic resistance genes. Consequently, accurate identification of plasmids from whole-genome sequencing (WGS) data remains one of the major challenges for sequencing-based precision medicine in infectious diseases. Here, we assess the heterogeneity of four state-of-the-art tools (cBar, PlasmidFinder, plasmidSPAdes and Recycler) for the *in silico* prediction of plasmid-derived sequences from WGS data. Heterogeneity, sensitivity and precision were evaluated by reference-independent and reference-dependent benchmarking using 846 Gram-negative clinical isolates. Interestingly, the majority of predicted sequences were tool-specific, resulting in a pronounced heterogeneity across tools for the reference-independent assessment. In the reference-dependent assessment, sensitivity and precision values were found to substantially vary between tools and across taxa, with cBar exhibiting the highest median sensitivity (87.45%) but a low median precision (27.05%). Furthermore, integrating the individual tools into an ensemble approach showed increased sensitivity (95.55%) while reducing the precision (25.62%). CBar and plasmidSPAdes exhibited the strongest concordance with respect to identified antibiotic resistance factors. Moreover, false-positive plasmid predictions typically contained only few antibiotic resistance factors. In conclusion, while high degrees of heterogeneity and variation in sensitivity and precision were observed across the different tools and taxa, existing tools are valuable for investigating the plasmid-borne resistome. Nevertheless, additional studies on representative clinical data sets will be necessary to translate *in silico* plasmid prediction approaches from research to clinical application.

Key words: bacteria; plasmids; prediction; next-generation sequencing

Introduction

Bacterial plasmids play important roles in the emergence and spread of antibiotic resistance [1]. These genetic elements vary in size, are mostly circular, can replicate independently and often encode resistance- and/or virulence-related genes [1–4]. Moreover, the dissemination of pathogens is facilitated by inter-species plasmid exchange [5]. A prominent example is the plasmid-encoded *mcr-1* gene inducing colistin resistance originally reported by Y. Liu and Y. Wang for Enterobacteriaceae samples collected in China [6]. The *mcr-1* gene was subsequently found in bacteria collected in Europe, Laos, Thailand and Nigeria [7]. Therefore, plasmid detection and classification are crucial steps for the identification and characterization of plasmid-mediated phenotypes.

Polymerase chain reaction-based replicon typing (based on elements of the replication machinery) [8, 9] and MOB typing (based on conserved motifs of the relaxase gene) are frequently used to detect and classify plasmids [10, 11]. Limitations of these approaches are, among others, that the available typing schemes do not cover all plasmids and that the complete genetic repertoire of the plasmid(s) remains unknown, as the focus of these approaches is on a specific set of genes [12]. In contrast, whole-genome sequencing (WGS) indiscriminately resolves the chromosomal and extrachromosomal genetic complements. Subsequent annotation of *de novo* assembled sequences enables the characterization of chromosome- and plasmid-derived functional potential in addition to taxonomic identification of the studied organism. In a detailed review of plasmid classification within the context of antibiotic resistance epidemiology, Orlek *et al.* [12] describe the potential of the *in silico* analysis of WGS data to address the limitations of replicon and MOB typing. Furthermore, Arredondo-Alonso *et al.* [13] reviewed computational solutions for the automated plasmid prediction on a set of 42 reference genomes. The existing *in silico* approaches can be divided into three main categories: marker-gene search-based approaches, e.g. searching for replicons in the sequences (PlasmidFinder [14]); approaches based on genomic signatures, e.g. *k*-mer frequencies, of plasmid-derived and chromosomal DNA (cBar [15]); and approaches identifying plasmids based on *k*-mer coverage differences and/or circular paths in the assembly graph (PlasmidSPAdes [16], Recycler [17]). However, repetitive regions and/or genes found on multiple genomic units (chromosomes and plasmids) challenge the *de novo* assembly of short-read sequencing data, resulting in fragmented assemblies and mis-assemblies [17]. In accordance with studies reporting on the improved contiguity of genome assemblies based on or augmented by long reads [18–21], Arredondo-Alonso *et al.* conclude that long-read sequencing data are expected to greatly assist in the resolution of chromosomal and extrachromosomal sequences. Unarguably, full-length genomic resolution is ultimately desirable, but despite advances in long-read sequencing, short-read-based approaches currently dominate the WGS space and can provide crucial diagnostic information. Therefore, the analysis of a cohort of clinical samples will allow improved assessment of the variance in the predictions across different taxa but also between the individual tools.

Here, we analyzed the short-read WGS data of 846 Gram-negative, clinical bacterial isolates using four existing *in silico* plasmid prediction tools (cBar, PlasmidFinder, plasmidSPAdes and Recycler) and an ensemble approach that integrates the individual tools' predictions. The heterogeneity between the individual tools was first assessed using reference-independent approaches. Subsequently, an *ad hoc* ground truth was defined. This was necessary as the herein included isolates were patient-derived and the closest reference genome needed to be identified first. *De novo* assembled contigs were then aligned against the respective reference chromosome(s) and plasmid(s) to identify plasmid-positive samples. This information was used to evaluate the sensitivity and precision of the individual tools and the ensemble approach. Furthermore, the differences in *k*-mer coverage of chromosome and plasmid sequences in plasmid-positive samples were compared. Finally, we analyzed the concordance between the predictions and the ground truth with respect to plasmid-borne antibiotic resistance genes.

Materials and methods

WGS and preprocessing

Batches of 96 samples were sequenced per lane for paired-end sequencing (2 × 100 bp) on Illumina Hiseq2000 or Hiseq2500 sequencers using TruSeq PE Cluster v3 and TruSeq SBS v3 sequencing chemistry (Illumina) as previously described in detail [22]. A total of 2 705 458 738 raw reads and a median of 2 987 123 reads per sample were generated. Trimmomatic version 0.35 was used with the command line parameters: 'PE ILLUMINACLIP:NexteraPE-PE.fa:1:50:30 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36' [23]. Only the trimmed, paired-end reads were used herein, if not stated otherwise.

De novo assembly

SPAdes version 3.10.1 [24] was used to assemble the trimmed, paired-end reads with the following parameters: '--careful -t 6 -k 21,33,55'.

Predicting plasmid sequences

plasmidSPAdes: 'plasmidspades.py' from SPAdes version 3.10.1 [16, 24] was used to assemble the trimmed, paired-end reads to identify candidate plasmid sequences, with the following parameters: '--careful -t 6 -k 21,33,55'.

PlasmidFinder: Sequences for the Enterobacteriaceae were downloaded from https://bitbucket.org/genomicepidemiology/plasmidfinder_db/src (commit ID: d5a49e9b01b0) [14]. A BLASTN database was built using 'makeblastdb' of ncbi-blast-2.6.0+.

cBar: Version 1.2 was used [15].

Recycler: Version 0.62 was used [17]. The required BAM file was generated using bwa-0.7.15. Recycler's 'make_fasta_from_fastg.py' was used to generate the FASTA file (from the 'assembly_graph. fastg' file generated by

SPAdes) required to build the bwa index [25, 26]. The trimmed paired-end reads were aligned against the resulting index with ‘bwa mem’, and the SAM output was directly converted to BAM format using ‘samtools view -buS -j samtools view -bF 0x0800 -j samtools sort –’ (samtools version 0.1.19-96b5f2294a) [27]. The resulting BAM file was indexed using ‘samtools index’. Finally, Recycler was run with the following options: ‘-g assembly_graph.fastg -k 55 -b assembly_graph.bam -i True’.

Ensemble approach: To increase the sensitivity, we implemented a straightforward ensemble approach. The candidate plasmid sequences, as predicted by cBar, plasmidSPAdes, PlasmidFinder and Recycler, were pooled and clustered using ‘cd-hit-est’ from CD-HIT version 4.6.6 and the default parameters [28, 29]. Accompanying code can be found at https://github.com/claczny/2017_plasmid_prediction_review.

Pairwise correlation of the individual tools’ predictions

Sourmash [30] version 2.0.0a1 was used to compute signatures for each tool’s predicted sequences (‘compute -k 31 –scaled 50 –track-abundance’). As plasmid-derived sequences are typically much shorter than chromosomal sequences, a small scaling factor was chosen accordingly. Subsequently, for all tools, all predictions within each tool were compared against each other using sourmash’s ‘compare’ function. The resulting similarity matrices per tool were converted to distance matrices (1—similarity value), and all pairwise tool combinations were correlated using the ‘mantel’ function (‘method=""pearson”, permutations=999, parallel=30’) in the vegan R package [31] for samples occurring in both of the predictions of the respective tool pair. The superheat-function in the superheat R package was used for plotting.

Complete reference genomes

Nucleotide FASTA files of complete bacterial genomes were downloaded from the NCBI RefSeq database (ncbi-genome-download, https://github.com/kblin/ncbi-genome-download, version 0.2.2, parameters: --section refseq --format fasta --assembly-level complete --human-readable --parallel 5 --retries 3 --verbose bacteria, on 24 May 2017). In total, 6901 genomes were retrieved; sequences containing the word ‘plasmid’ in their FASTA header were considered as plasmids resulting in 5611 plasmid and 7415 non-plasmid sequences in total.

Defining the *ad hoc* ground truth data

Lacking dedicated, finished genomes for the present clinical, i.e. patient-derived in contrast to reference material, isolates, sourmash version 2.0.0a1 was used to identify the most similar, complete reference genome. Specifically, signatures were first computed for each complete reference genome and the contigs of each successful *de novo* assembly (‘compute -k 31 --scaled 2000 --track-abundance -o SEQ.sig SEQ.fa’). The reference genomes’ signatures were indexed (‘index REFIDXPREFIX -k 31 --traverse-directory PATH_TO_REF_SIGNATURES’). Subsequently, for each *de novo* assembly, the index was searched (‘search -k 31 ASSEMBLY.sig REFIDXPREFIX.sbt.json -o ASSEMBLY.best_only_hits.txt --best-only’), and the top hit returned by sourmash was used as the respective reference. For each isolate-reference pair, the isolate’s *de novo* contigs were aligned against the reference to identify plasmid or chromosome sequences using BLASTN (ncbi-blast-2.6.0+, format: ‘6 std qcovs qcovhsp qlen slen’) [32]. For each query sequence, the subject (reference sequence) with the longest alignment length and highest query-coverage-by-subject was selected. If multiple hits existed, the hit with the highest bit-score was chosen. If multiple hits remained, the first subject representing a plasmid was chosen. Thus, each *de novo* contig was assigned a label whether it represents a plasmid, and contigs not matching a sequence of the closest reference genome were considered ‘unclassified’.

Evaluating the predictions

Reference-independent analysis of heterogeneity: Sequences were clustered as described for the ‘Ensemble approach’. The ‘clstr2txt.pl’ script from CD-HIT version 4.6.6 was used to reformat the cluster output. The reformatted output was used to compute the fraction of the cumulative length of the cluster centroids that was represented by one, two, three or all four tools. It should be noted that the length of the cluster centroid was used here as a proxy. However, the actual shared fraction could be lower if cluster members are of shorter length than the cluster centroid.

Reference sequence coverage: PlasmidSPAdes and Recycler generate their own set of contigs, whereas cBar and PlasmidFinder directly identify candidate plasmid sequences on the *de novo* assembled contigs. Thus, to stay consistent between all tools, the predicted sequences were linked with the *ad hoc* ground truth sequences by using the former as queries and the latter as the subjects in BLASTN (ncbi-blast-2.6.0+, format: ‘6 std qcovs qcovhsp qlen slen’). Similar to the approach used for defining the ground truth data, for each query sequence, the subject (*de novo* contig) with the longest alignment length and highest query-coverage-by-subject was selected. If multiple hits existed, the hit with the highest bitscore was chosen. If multiple hits remained, the longest subject was chosen. Should still multiple hits remain, the first subject representing a plasmid was chosen. Unclassified sequences were ignored. Based on the resulting prediction-to-ground-truth mapping, the sensitivity and precision were computed using the following definitions:

- P = cumulative length of ground truth plasmid sequences
- $TP = \sum length(subject)$; if query was predicted plasmid and subject was ground truth plasmid
- $FN = P - TP$
- N = cumulative length of ground truth chromosome sequences
- $FP = \sum length(subject)$; if query was predicted plasmid and subject was ground truth chromosome
- $TN = N - FP$
- $Sensitivity = TP / P$
- $Precision = TP / (TP + FP)$

The following edge cases were considered and handled accordingly:

- The sample contained no plasmids and the tool predicted no plasmids: $P = 0$, $TP = 0$, $FN = 0$, $FP = 0$, $TN = N$
- The sample contained plasmids and the tool predicted no plasmids: $TP = 0$, $FN = P$, $FP = 0$, $TN = N$
- The sample contained no plasmids and the tool predicted plasmids: $P = 0$, $TP = 0$, $FN = 0$

Antibiotic resistance genes: Prokka version 1.11 was used to annotate the genes of the predicted and ground truth plasmid sequences [33]. Translated coding DNA sequences were searched against the ResFams core database using hmmsearch version 3.1b2 ('--cut_tc --tblout') [34, 35]. The counts of ResFams hits per-sample-and-tool were compared with the respective counts of the ground truth for samples common to the respective tool and the ground truth. The Spearman correlation was computed using the 'cor' function in R version 3.3.2 [36]. For each comparison, a linear model including confidence intervals was fitted using the 'geom_smooth' function from the ggplot package version 2.2.1 in R version 3.3.2 [37].

Results and discussion

Cultured isolates of Gram-negative bacteria from 846 clinical samples were sequenced as described in Galata *et al.* [22], and *de novo* assemblies were successfully created for 844 samples. We evaluated the performance of four plasmid-prediction tools: cBar, PlasmidFinder, plasmidSPAdes and Recycler. Moreover, we integrated the individual tools into an ensemble approach by merging and clustering the predictions according to their nucleotide sequence identity to remove redundant sequences. In addition to evaluating the predictions using reference-independent as well as reference-dependent approaches, the concordance between the predictions and the ground truth with respect to plasmid-borne antibiotic resistance genes was analyzed.

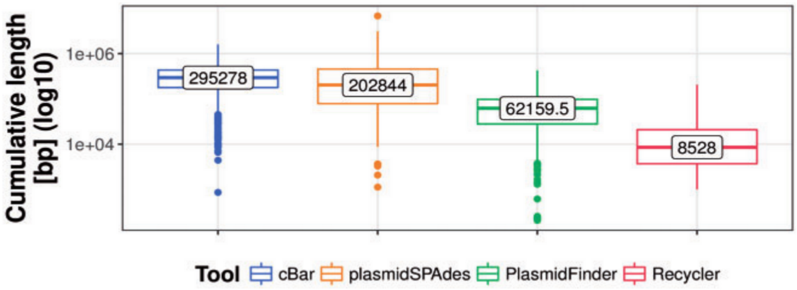


Figure 1. Cumulative lengths of the predicted plasmid sequences per tool. The y-axis uses a log10 scale. The median values are shown and the boxplots represent the median, two hinges, two whiskers and all outlier points individually.

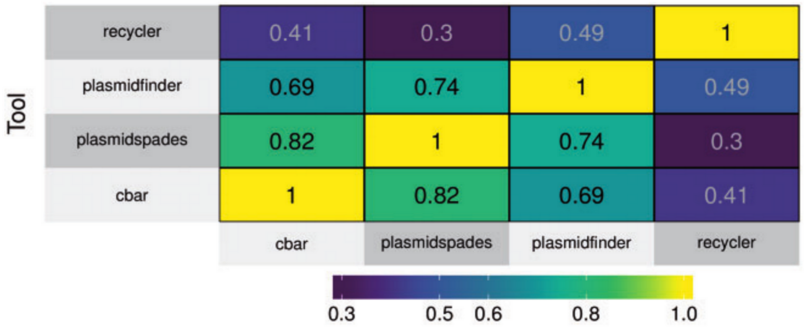


Figure 2. Correlation of the tools' predictions. For each tool, a distance matrix with respect to the tool's predictions was computed. Pairwise distance matrix correlation was computed and is shown in the heatmap. The color indicates the correlation degree and correlation values are shown in each cell.

Reference-independent assessment of plasmid predictions

In a first analysis, we were mainly interested in the heterogeneity of the predictions between the individual tools. We thus performed the predictions and compared them with each other. Interestingly, all tested tools substantially varied in their number of predicted plasmid-positive samples, i.e. samples predicted to contain at least one plasmid-derived sequence. CBar predicted all 844 samples to be plasmid-positive, while plasmidSPAdes, PlasmidFinder and Recycler predicted 766, 446 and 375 plasmid-positive samples, respectively. Moreover, the cumulative lengths of the predicted sequences per sample were found to vary markedly: cBar was found to have the largest cumulative lengths, while Recycler had the lowest (Figure 1).

Moreover, the tools were tested for their pairwise correlations across the predictions. CBar and plasmidSPAdes were found to exhibit the highest correlation value (0.82), suggesting that these two approaches resulted in somewhat related predictions (Figure 2). In contrast, plasmidSPAdes and Recycler were found to have the lowest pairwise correlation (0.3). Based on the clustering of the individual tools' predictions using the ensemble approach, the tools' heterogeneity was furthermore evaluated with respect to the shared fraction of the cumulative plasmid lengths. The largest fraction was represented by sequences predicted by a single tool (Figure 3). Conversely, all four tools were infrequently found to show pronounced overlap in their prediction. Furthermore, strong variations were observed in the fraction of cumulative length predicted by one or by two tools. This indicates a distinct heterogeneity between the individual tools' predictions.

Reference-dependent assessment of sensitivity and precision

A complete reference genome could be identified for 818 of the 846 samples. The *ad hoc* definition of the ground truth was required because of the lack of dedicated, finished genomes, e.g. using complimentary long-read sequencing data, for the present set of clinical isolates. The median cumulative lengths of chromosome contigs, plasmid contigs and unclassified contigs were 4 907 449, 114 954 and 27 733 bp, respectively (Supplementary Figure S1). Seven samples had >1 Mbp of unclassified contigs and were thus excluded from further analyses, resulting in median lengths of 514.0, 437.5 and 118.0 bp, for chromosome contigs, plasmid contigs and unclassified contigs, respectively (Supplementary Figure S2). A total of 347 samples were considered to be plasmid-positive according to the *ad hoc* ground truth and were subsequently used to compute the sensitivity and precision of the individual tools and of the ensemble approach (Supplementary Figure S3).

CBar was found to be the most sensitive (median sensitivity: 87.45%) among the individual tools, followed by plasmidSPAdes (81.49%) and PlasmidFinder (36.47%) (Figure 4). Recycler's predictions generally had overall low cumulative lengths (Figure 1), consequently resulting in extremely low sensitivity values (median sensitivity: 0.00%). Importantly, Recycler was designed to recover circular sequences, and the present results suggest that their number was minimal in our *de novo* assemblies. The ensemble approach resulted in a median sensitivity value of 95.55%.

Resolving the prediction performances by genus revealed strong variations, both within and between the individual tools (Supplementary Figure S4). For example, while cBar was found to exhibit overall high sensitivity values, plasmid sequences of *Acinetobacter* spp. were less well detected. Moreover, the sensitivity of plasmidSPAdes varied strongly for the *Citrobacter* spp., *Enterobacter* spp. and *Salmonella* spp. samples. PlasmidFinder exhibited particularly low sensitivity for *Acinetobacter* spp., which is to be expected, as this genus is a member of the Moraxellaceae family and, thus, not covered by PlasmidFinder's Enterobactericeae-specific database. The sensitivity of the ensemble approach was found to be on par or better compared with the individual tools.

While cBar had the highest median sensitivity, its median precision (27.05%) was below the median precision of PlasmidFinder (100%) and plasmidSPAdes (52.70%), indicating that cBar frequently misclassifies chromosomal contigs as being plasmid-derived (false positives). The median precision of the ensemble approach was 25.62%. Importantly, the ensemble approach included all the false-positive predictions of the individual tools, which explains the low precision. Similar to the sensitivity results resolved by genus, the precision of the individual tools varied substantially (Supplementary Figure S4). Notably, the highest median precisions were observed for *Klebsiella* spp. In contrast, the precision was extremely low for *Acinetobacter* spp., regardless of the approach being reference-dependent (cBar, PlasmidFinder) or reference-independent (plasmidSPAdes, Recycler). Hierarchical clustering of the individual tools and the ensemble approach with respect to their true-positive values revealed cBar and the ensemble approach to be the most similar, followed by plasmidSPAdes (Figure 5).

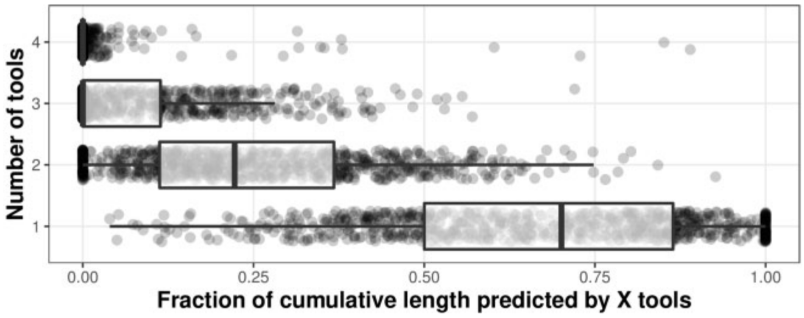


Figure 3. Fraction of cumulative lengths shared by the tested tools. The fraction of the cumulative length is shown on the x-axis, and the number of tools exhibiting overlap of the respective sequence(s) is shown on the y-axis. The lengths of the cluster centroids were taken as proxies. Points are jittered randomly vertically per tool for representation purposes. The boxplots represent the median, two hinges and two whiskers.

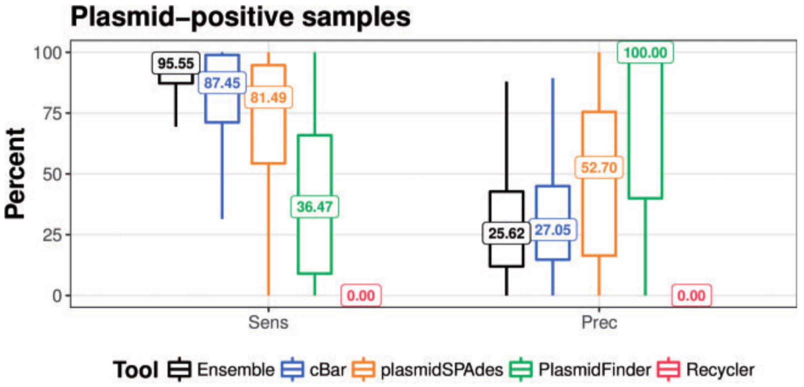


Figure 4. Prediction performances of the tested tools and the ensemble approach for plasmid-positive samples based on the *ad hoc* ground truth. Sensitivity ('Sens') and precision ('Prec') are shown. The median values are shown and the boxplots represent the median, two hinges and two whiskers.

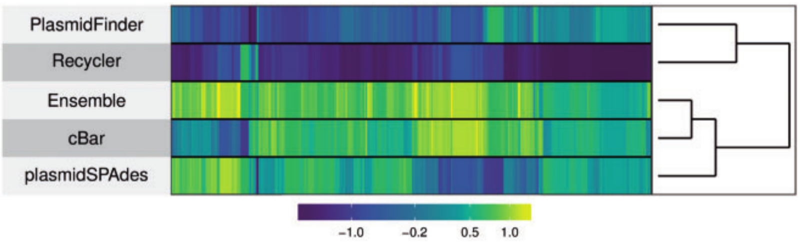


Figure 5. Hierarchical clustering of individual tools according to their true-positive values. True-positive values represent the cumulative base pair length correctly covered by the individual tools and were scaled and centered before computing the hierarchical clustering.

Differential coverage between chromosome and plasmid sequences

Plasmids can independently replicate [3, 4] and thus can occur in different copy numbers than the bacterial chromosome(s). PlasmidSPAdes uses this information to identify assembly graph components with (substantially) differing coverage, considering these components as candidate plasmids. However, this approach is, by design, challenged by plasmids occurring in similar copy numbers as the chromosome(s) (false negatives), or by components within the graph that exhibit coverage differences despite representing chromosomal sequences (false positives), e.g. because of bacterial cells at different stages in the replication cycle [38]. To study how frequently plasmid sequences significantly differed in their copy numbers from the chromosome sequences, we analyzed the *k*-mer coverage of the *de novo* assembled contigs. Of the 811 isolates (818 – 7 samples with >1 Mbp of unclassified sequences), 28.11% (228 of 811) showed statistically significant results (alpha = 0.05; false discovery rate-adjusted: 185 of 811) when tested for unimodality of the *k*-mer coverage distributions (Supplementary Figure S5), suggesting that these distributions could be considered mutimodal. However, only 31.70% (110 of 347) of the plasmid-positive samples were likely multimodal in their *k*-mer coverage distributions. Moreover, 61.10% (212 of 347) of the plasmid-positive samples significantly differed in their *k*-mer coverages of the plasmid sequences and chromosome sequences (Wilcoxon-Mann-Whitney-test, $P < 0.05$). It should be noted that plasmidSPAdes median sensitivity was higher (81.29%; Figure 4); yet, this was computed using the sequence coverage of plasmid sequences rather than number of samples. Correctly predicted, long-assembled sequences will increase the true-positive value, thereby leading to higher sensitivity values.

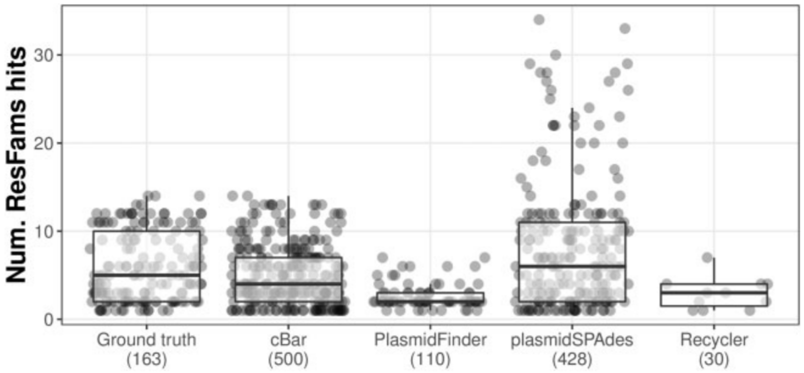


Figure 6. ResFams hits counts of plasmid-positive samples. The number of samples per tool is shown in parentheses. Only plasmid-positive samples with at least one ResFams hit are shown. Points are jittered randomly horizontally per tool for representation purposes. The boxplots represent the median, two hinges and two whiskers.

Antibiotic resistance factors encoded in plasmid sequences

The *in silico* separation of genomic sequences into 'chromosome-derived' or 'extrachromosome-derived' has proven to be a challenging task as demonstrated herein as well as in [12, 13]. Nevertheless, the identification of candidate plasmid-derived sequences in fragmented assemblies is relevant. Specifically, the functional potential can thus be assessed for the candidates. To this end, antibiotic resistance genes included in ResFams were identified on the predicted and ground truth plasmid sequences of plasmid-positive samples. The number of ResFams hits was found to vary within and between the individual tools but also for the ground truth (Figure 6). PlasmidFinder and Recycler recovered few of the expected ResFams hits, which is in accordance with the reduced sensitivity observed herein (Figure 4). CBar and plasmidSPAdes were found to more closely represent the ground truth distribution of the ResFams hits. Only plasmidSPAdes exhibited a higher number of hits than found in the ground truth. These extra hits might represent chromosome-borne antibiotic resistance genes. As plasmidSPAdes uses coverage information for its predictions, it could be speculated that the respective chromosomal regions exhibited differential coverage to the remainder of the chromosome. While there are various potential reasons as to why this could occur, e.g. competitive advantage under antibiotic pressure and thus increased replication, the exact reason is currently unknown. Moreover, the ResFams hit counts were compared pairwise between the ground truth and the individual tools, and the respective Spearman correlations were computed (Figure 7). CBar and plasmidSPAdes were found to be the closest to represent the ground truth, with cBar exhibiting a higher correlation (0.68 versus 0.56), likely because of the increased variation toward low or high counts for plasmidSPAdes.

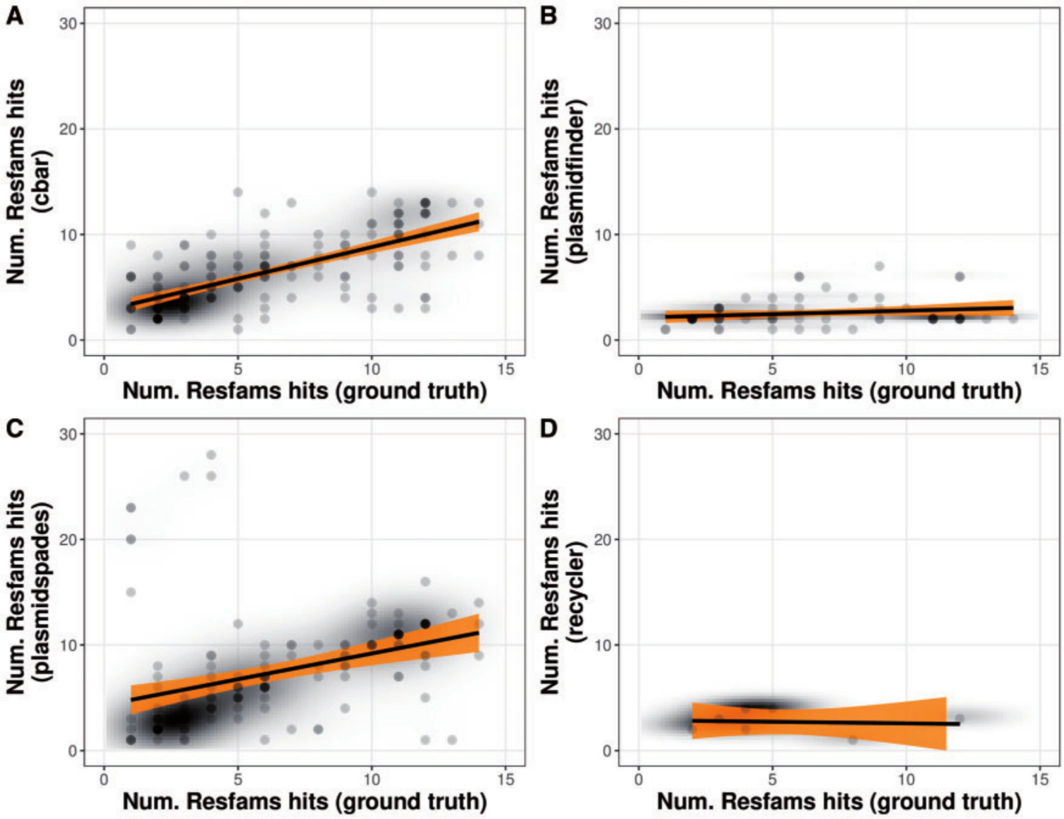


Figure 7. Comparison of ResFams hits against the *ad hoc* ground truth. ResFams hits counts of the four herein tested tools are plotted against the respective counts in the ground truth for paired samples. A linear model was fitted (black line) and confidence intervals are shown (in orange). Moreover, a two-dimensional density estimate is plotted with transparency increasing with decreasing point density.

Conclusion

The importance of WGS has been repeatedly demonstrated for taxonomic identification of microorganisms, with its application in infectious disease diagnostics and in epidemiological studies providing direct benefits to individuals and the general public [39–41]. Furthermore, the indiscriminate extraction of the entire microbial genomic complement enables concurrent sequencing of chromosomal and extrachromosomal sequences, e.g. plasmids in bacteria. This is especially relevant as plasmid-encoded functions can strongly affect the bacterial phenotype, thus providing crucial information beyond chromosomes and taxonomy [42–45]. To this end, the present study analyzed the performances of four plasmid prediction tools on *de novo* assemblies of 846 Gram-negative WGS clinical isolates using reference-independent and reference-dependent evaluation approaches. With respect to the latter, the use of patient-derived isolates, in contrast to reference material, required the definition of an *ad hoc* ground truth. This approach was found to be robust for the plasmid-positive samples, as the cumulative length of unclassified sequences was limited. However, plasmid sequences, in particular if they were recently acquired, might have been missed; yet, this would not negatively affect the present evaluation, as unclassified sequences were ignored. Moreover, plasmid sequences recently introduced in the chromosome(s) or plasmid sequences homologous to chromosome sequences might represent confounding factors in the definition of the *ad hoc* ground truth. This further highlights the importance of full-length assemblies/reference genomes, which were, however, unavailable for the herein included isolates, and the generation of this complementary data was beyond the scope of the current study.

Overall, no single-best approach was identified and pronounced variations in heterogeneity between the tools were observed, with cBar and plasmidSPAdes showing the strongest correlation. Moreover, the diversity of the present samples comprised 11 genera of at least 20 samples and allowed to reveal taxon-dependent variation, both, within tools and between tools. Interestingly, *Acinetobacter*-borne plasmids were less well detected by cBar, resulting in a low sensitivity, which may be because of a limited representation of this genus in the reference database that was originally used for cBar's training. Furthermore, the generally low precision for this specific genus suggests that *Acinetobacter* spp. infections may require dedicated analyses and attention, e.g. in the case of plasmid-carrying, multidrug-resistant *Acinetobacter baumannii* organisms [46–48]. The taxon-dependent variation in the tools' performances highlights the importance of concurrent identification of taxonomy and functional potential and the need for reference databases with an increased diversity, e.g. improved coverage of *Acinetobacter* spp. by cBar and PlasmidFinder. Moreover, we showed that copy numbers of plasmid sequences need not necessarily vary significantly to the copy number of the chromosome(s), thereby limiting coverage-based approaches. Accordingly, the use of complementary approaches that could lend mutual support, e.g. using cBar and plasmidSPAdes, appears sensible.

In addition to the individual tools, an ensemble approach integrating the four independent predictions was evaluated. Overall, the sensitivity was found to be increased and less variable. However, the combination of the individual tools also led to reduced precision. Accordingly, the ensemble approach represents an interesting solution if the objective is to maximize the sensitivity, and false positives are acceptable and/or can be removed downstream, e.g. by identifying sequences with exceptionally high or low fold-coverage or by identifying sequences encoding relevant factors, such as antibiotic resistance genes. This approach is, however, not intended to replace the development of improved databases and prediction algorithms in the future. An example of the fast developments in this field is PlasmidTron, which was published, while the present manuscript was in revision [49]. Moreover, PLACNET represents a recently published approach for the plasmid reconstruction from WGS data [50]. It was excluded from the present evaluation of fully automated tools because of a manual pruning step in PLACNET's workflow.

The reconstructed genomic sequences, including the plasmid sequences, remained fragmented in the present study, which is in accordance with the results reported by Arredondo-Alonso *et al.* [13]. While long-read-based sequencing greatly improves the contiguity of genome assemblies [18, 19, 51], plasmid prediction tools can strongly reduce the search space for short-read-based data. Importantly, despite the frequent prediction of false positives, the accordance in the number of antibiotic resistance genes with respect to the ground truth was found to be high for cBar and plasmidSPAdes. Overall, this is expected to support precision medicine by reducing the time and work burden required for data examination. Furthermore, the present study illustrates that specific objectives are met by specific approaches and, thus, systematic benchmarking on extensive and curated data sets is important for the translation of bioinformatics tools from research to clinical application.

Key Points

- Extrachromosomal DNA in the form of plasmids can carry phenotype-relevant information, e.g. antibiotic resistance factors.
- Next-generation sequencing of isolates allows linkage of taxonomy and extrachromosomal functional potential via concurrent resolution of chromosomal and extrachromosomal DNA.
- Existing *in silico* plasmid-prediction approaches showed limited agreement as well as strong inter- and intra-taxon variability on a set of 846 WGS clinical bacterial isolates.
- Combining the individual predictions resulted in increased sensitivity while reducing precision.
- Antibiotic resistance gene counts on predicted plasmid sequences were not strongly affected by false-positive predictions.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgement

The authors would like to thank Curetis GmbH and Ares Genetics GmbH for the support as well as the provided data set.

Funding

In parts by the Best Ageing (grant number 306031) from the European Union.

Availability of WGS data

The raw WGS data are available on a reasonable request for academic research use only after signing a data transfer agreement.

References

1. Carattoli A. Resistance plasmid families in Enterobacteriaceae. Antimicrob Agents Chemother 2009;53(6):2227–38.

2. Frost LS, Leplae R, Summers AO, et al. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 2005;3(9):722–32.

3. Scott JR. Regulation of plasmid replication. Microbiol Rev 1984; 48(1):1–23.

4. del Solar G, Giraldo R, Ruiz-Echevarría MJ, et al. Replication and control of circular bacterial plasmids. Microbiol Mol Biol Rev 1998;62(2):434–64.

5. Conlan S, Park M, Deming C, et al. Plasmid dynamics in KPC-positive Klebsiella pneumoniae during long-term patient colonization. MBio 2016:2:e000085.

6. Liu Y-Y, Wang Y, Walsh TR, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis 2016;16:161–8.

7. Zhi C, Lv L, Yu L-F, et al. Dissemination of the mcr-1 colistin resistance gene. Lancet Infect Dis 2016;16:292–3.

8. Couturier M, Bex F, Bergquist PL, et al. Identification and classification of bacterial plasmids. Microbiol Rev 1988;52(3): 375–95.

9. Carattoli A, Bertini A, Villa L, et al. Identification of plasmids by PCR-based replicon typing. J Microbiol Methods 2005;63(3): 219–28.

10. Francia MV, Varsaki A, Garcillán-Barcia MP, et al. A classification scheme for mobilization regions of bacterial plasmids. FEMS Microbiol Rev 2004;28(1):79–100.

11. Alvarado A, Garcillán-Barcia MP, de la Cruz F. A degenerate primer MOB typing (DPMT) method to classify gamma-proteobacterial plasmids in clinical and environmental settings. PLoS One 2012;7(7):e40438.

12. Orlek A, Stoesser N, Anjum MF, et al. Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. Front Microbiol 2017;8:182.

13. Arredondo-Alonso S, Willems RJ, van Schaik W, et al. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microb Genomics 2017; 1–18.

14. Carattoli A, Zankari E, García-Fernández A, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 2014;58(7):3895–903.

15. Zhou F, Xu Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics 2010;26(16): 2051–2.

16. Antipov D, Hartwick N, Shen M, et al. plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics 2016;32:3380–7.

17. Rozov R, Brown Kav A, Bogumil D, et al. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. Bioinformatics 2016;33:475–82.

18. English AC, Richards S, Han Y, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 2012;7(11):e47768.

19. Reuter S, Hunt M, Peacock SJ, et al. Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. Microb Genomics 2016;2(9): e000085.

20. George S, Pankhurst L, Hubbard A, et al. Resolving

plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. Microb Genomics 2017;3(8):10.

21. Wick RR, Judd LM, Gorrie CL, et al. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017;13(6):e1005595.

22. Galata V, Backes C, Laczny CC, et al. Comparing genome versus proteome-based identification of clinical bacterial isolates. Brief Bioinform 2016; doi:10.1093/bib/bbw122.

23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15): 2114–20.

24. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19(5):455–77.

25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14): 1754–60.

26. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN] 2013;1303.3997v1.

27. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078–9.

28. Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28(23): 3150–2.

29. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13):1658–9.

30. Titus Brown C, Irber L. sourmash: a library for MinHash sketching of DNA. J Open Source Softw 2016;1(5):27.

31. Oksanen J, Blanchet FG, Friendly M, et al. vegan: Community Ecology Package, 2017.

32. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;215(3):403–10.

33. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30(14):2068–9.

34. Eddy SR. Profile hidden Markov models. Bioinformatics 1998; 14(9):755–63.

35. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. Isme J 2015;9(1):207–16.

36. R Core Team. R: A Language and Environment for Statistical Computing, 2016.

37. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag, 2009.

38. Korem T, Zeevi D, Suez J, et al. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science 2015;349(6252):1101–6.

39. Grumaz S, Stevens P, Grumaz C, et al. Next-generation sequencing diagnostics of bacteremia in septic patients. Genome Med 2016;8(1):73.

40. Loman NJ, Constantinidou C, Christner M, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104: H4. Jama 2013;309(14):1502–10.

41. Zhou K, Lokate M, Deurenberg RH, et al. Characterization of a CTX-M-15 producing Klebsiella pneumoniae outbreak strain assigned to a novel sequence type (1427). Front Microbiol 2015; 6:1250.

42. von Wright A, Tynkkynen S. Construction of Streptococcus lactis subsp. lactis strains with a single plasmid associated with mucoid phenotype. Appl Environ Microbiol 1987;53(6): 1385–6.

43. Matsui H, Bacot CM, Garlington WA, et al. Virulence plasmid-borne spvB and spvC genes can replace the 90-kilobase plasmid in conferring virulence to Salmonella enterica serovar typhimurium in subcutaneously inoculated mice. J Bacteriol 2001; 183(15):4652–8.

44. Hammerl JA, Freytag B, Lanka E, et al. The pYV virulence plasmids of Yersinia pseudotuberculosis and Y. pestis contain a conserved DNA region responsible for the mobilization by the self-transmissible plasmid pYE854. Environ Microbiol Rep 2012; 4(4):433–8.

45. Guiney DG, Fang FC, Krause M, et al. Plasmid-mediated virulence genes in non-typhoid Salmonella serovars. FEMS Microbiol Lett 1994;124(1):1–9.

46. Huang H, Dong Y, Yang Z-L, et al. Complete sequence of pABTJ2, a plasmid from Acinetobacter baumannii MDR-TJ, carrying many phage-like elements. Genomics Proteomics Bioinformatics 2014;12(4):172–7.

47. Weber BS, Ly PM, Irwin JN, et al. A multidrug resistance plasmid contains the molecular switch for type VI secretion in Acinetobacter baumannii. Proc Natl Acad Sci USA 2015;112(30): 9442–7.

48. Hamidian M, Holt KE, Pickard D, et al. A small Acinetobacter plasmid carrying the tet39 tetracycline resistance determinant. J Antimicrob Chemother 2016;71(1):269–71.

49. Page AJ, Wailan A, Shao Y, et al. PlasmidTron: assembling the cause of phenotypes from NGS data. bioRxiv 2017. https://doi.org/10.1101/188920.

50. Lanza VF, de Toro M, Garcillán-Barcia MP, et al. Plasmid flux in Escherichia coli ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. PLoS Genet 2014;10:e1004766.

51. Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION TM portable single-molecule nanopore sequencer. Gigascience 2014;3(1):22.

3.5 *PLSDB: a resource of complete bacterial plasmids*

Published online 31 October 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D195–D202***
doi: 10.1093/nar/gky1050

PLSDB: a resource of complete bacterial plasmids

Valentina Galata[*], Tobias Fehlmann, Christina Backes and Andreas Keller[*]

Chair for Clinical Bioinformatics, Saarland University, Campus Building E2.1, 66123 Saarbruecken, Germany

Received August 10, 2018; Revised October 15, 2018; Editorial Decision October 16, 2018; Accepted October 17, 2018

ABSTRACT

The study of bacterial isolates or communities requires the analysis of the therein included plasmids in order to provide an extensive characterization of the organisms. Plasmids harboring resistance and virulence factors are of especial interest as they contribute to the dissemination of antibiotic resistance. As the number of newly sequenced bacterial genomes is growing a comprehensive resource is required which will allow to browse and filter the available plasmids, and to perform sequence analyses. Here, we present PLSDB, a resource containing 13 789 plasmid records collected from the NCBI nucleotide database. The web server provides an interactive view of all obtained plasmids with additional meta information such as sequence characteristics, sample-related information and taxonomy. Moreover, nucleotide sequence data can be uploaded to search for short nucleotide sequences (e.g. specific genes) in the plasmids, to compare a given plasmid to the records in the collection or to determine whether a sample contains one or multiple of the known plasmids (containment analysis). The resource is freely accessible under https://ccb-microbe.cs.uni-saarland.de/plsdb/.

INTRODUCTION

Naturally occurring bacterial plasmids are a key consideration when studying bacterial isolates or communities as they can contain genes giving their host an adaption advantage (1). In the context of bacterial pathogens, antibiotic resistance and virulence genes located on these extra-chromosomal DNA molecules are of particular interest. As the plasmids can be exchanged between bacterial cells, the knowledge about their distribution is crucial to study the spread of plasmids harboring relevant genetic markers (2). Thus, newly sequenced plasmids need to be compared to the already published sequences to determine whether they have been already detected in other organisms. The importance of tracking clinically relevant plasmid or gene sequences was recently demonstrated after the discovery of the first plasmid-mediated resistance mechanism against colistin (MCR-1) in Enterobacteriaceae (3).

As the number of sequenced plasmids grows constantly together with the number of sequenced bacterial genomes and metagenomes (4), there is a need for a comprehensive overview of the already discovered plasmids providing information on their characteristics and distribution among different organisms. Though, NCBI already provides a list of plasmids from the RefSeq database (https://www.ncbi.nlm.nih.gov/genome/plasmids/) further utilities for the analysis using only this subset of records are currently not available. For example, the table can only be sorted but not filtered or searched, there is no information on associated samples and assemblies, and there is no BLAST database option available to search in these plasmid records only. Moreover, some of the NCBI records tagged as plasmids are mislabeled chromosomal sequences and many entries do not represent complete records making a filtering of these entries challenging (4). At the same time, the number of alternative plasmid resources is limited. The Addgene Repository stores plasmids used in the lab (5) and thus does not primarily focus on naturally occurring bacterial plasmids. The Plasmid Genome Database (PGD) was published as a resource of all fully sequenced plasmids (6); however, it seems that it is not maintained anymore as it is not accessible (http://www.genomics.ceh.ac.uk/plasmiddb/, accessed on 7 August 2018). Orlek *et al.* created a dataset of complete plasmids collected from the NCBI nucleotide database but it is limited to records from the family *Enterobacteriaceae* (7). Another dataset of finished bacterial plasmids was created by Robertson and Nash to be used as reference data in a software suit for processing plasmids from draft assemblies (8). A much more comprehensive collection of bacterial plasmids among the herein listed resources is offered by the recently launched web server pATLAS (http://www.patlas.site/). But, this resource does currently not allow for sequences to be uploaded and searched against the plasmids in the database; only the results obtained using the pATLASflow pipeline can be submitted (https://github.com/tiagofilipe12/pATLASflow, accessed on 1 August 2018).

To this end, we implemented a resource, PLSDB, including an extensive set of complete bacterial plasmids from the NCBI database covering records from RefSeq

[*]To whom correspondence should be addressed. Tel: +49 681 302 68612; Fax: +49 681 302 58094; Email: valentina.galata@uni-saarland.de
Correspondence may also be addressed to Andreas Keller. Tel: +49 681 302 68611; Fax: +49 681 302 58094; Email: andreas.keller@ccb.uni-saarland.de

and INSDC (which includes DDBJ, EMBL-EBI and GenBank). The plasmid records were annotated using ARG-ANNOT (9), CARD (10), ResFinder (11) and VFDB (12), and characterized by PlasmidFinder and pMLST (13). Also, additional metadata such as taxonomy, sequence features and sample information was incorporated. The database provides a user-friendly and interactive overview of the plasmid sequences which can be filtered and searched by various parameters. It also offers an option to search for short nucleotide sequences (e.g. genes) using BLASTn (14), to compare a plasmid sample represented by one or multiple nucleotide sequences to all included plasmids using Mash (15) and to perform a containment analysis (16), i.e. the identification of plasmids present within a sample representing a mixture of chromosome- and/or plasmid-derived sequences (https://genomeinformatics.github.io/mash-screen). The user can upload the sequence data to the web server or download the required BLAST database and Mash sketch files to run the analysis locally for batch analyses. We describe how PLSDB can be used for the analysis of sequencing data and compare our resource to the existing alternatives listed above.

PLASMID COLLECTION

All plasmid records were collected from the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore) from the resources INSDC (which includes DDBJ, EMBL-EBI and GenBank) and RefSeq using command line utilities EDirect (17) (version 9.80). The herein described data were retrieved on 14 September 2018.

Data retrieval and processing pipeline

Data collection. Plasmid records were searched in the NCBI nucleotide database by using the query from Orlek *et al.* (4) and filtering the results to have 'plasmid' as location tag, being assigned to a bacterial organism and being from the specified resource (INSDC or RefSeq). Document summary was fetched for each hit and the following information was extracted if available: UID, caption (accession without the version number), title (sequence description), creation date, topology (e.g. circular or linear), completeness, taxon ID, genome tag and sequence length. For the record taxon IDs, the associated name and rank, the complete lineage and the taxon ID and name for the ranks species, genus, family, order, class, phylum and superkingdom were obtained. For each BioSample ID linked to a plasmid record, the location name and coordinates, and the isolation source were extracted. The retrieved location coordinates were processed and if these were not available the location name was queried using the API of OpenCageData (https://opencagedata.com/). In the latter case, the mapped coordinates were manually checked to correct assignments deviating significantly from the expected location (e.g. wrong continent or country). For each assembly ID linked to a plasmid record, its completeness status, sequence release and submission date were extracted, and whether it is the latest assembly version. If a plasmid record was linked to multiple assembly IDs only the assembly with the tag 'latest' was assigned to this record. If none of the linked assemblies had this tag the newest one was chosen based on the sequence release date.

Record filtering. Subsequently, the collected plasmid records were filtered in several steps to remove incomplete or mislabeled chromosomal sequences. First, the plasmid records were filtered by their description using the regular expression defined by Orlek *et al.* (4), by their completeness and assembly completeness tags, and by their taxonomy to remove non-bacterial sequences. The record was required to have the completeness tag 'complete' and its assembly the tag 'Complete Genome'; if no assembly was associated with the record then only the record tag was used and *vice versa*; empty completeness tags were ignored, i.e. only the non-empty ones were used to remove the records. In the second step, the records were deduplicated: pairs of likely equal records were created using Mash (15) by computing the sketches of the plasmid sequences and their pair-wise distances. The sequences of pairs with a distance of zero were compared and identical records were grouped together. For each group, one record was chosen, similar to the approach described by Orlek *et al.* (4), by preferring RefSeq records over the INSDC records and by preferring records with additional information (mapped location coordinates and having a linked assembly). In ambiguous cases, the record with the older creation date was chosen. In the third filtering step, putative chromosomal sequences were identified and removed. A list of candidates was created by performing an *in silico* rMLST analysis (18), i.e. searching the 53 *rps* genes, downloaded from PubMLST (19) (https://pubmlst.org/rmlst/, 14 September 2018), in the plasmid records using BLASTn (14) (version 2.7.1+). The advantage of these markers for the detection of putative chromosomal sequences is their presence in all bacteria, their distribution around the chromosome, and their functional conservation (18). For the BLAST hits, the subject coverage was computed as $100 \cdot (\textit{alignment length} - \textit{total number of gaps})/\textit{subject length}$ and only hits with 100% identity and subject coverage were kept. As in some cases the *rps* genes can also be located on plasmids (20), only plasmid records having hits to more than 5 unique *rps* genes (i.e. more than 10% of the 53 genes) were subjected to a remote BLAST search (megablast) in the NCBI nr/nt database using an Entrez query to exclude non-chromosomal subject sequences. Any record having at least one hit with at least 99% identity and 80% query coverage was excluded from the plasmid collection.

Record annotation. The sequences were annotated by performing a BLASTn search for resistance factors from ARG-ANNOT (9), CARD (10) and ResFinder (11) with minimal identity and coverage of 95%, virulence factors from VFDB (12) with minimal identity and coverage of 95%, and replicons from PlasmidFinder (13) using the *Enterobacteriaceae* and the Gram-positive datasets with minimal identity of 80% and minimal coverage of 60%. For PlasmidFinder, the identity and coverage cutoffs were set according to authors' recommendations (13). The tool ABRicate, implemented by Seemann (https://github.com/tseemann/abricate, version 0.8.7), was used to download and prepare the databases which was done on 14 September 2018, except for VFDB which was updated on 17 September 2018. For the sequence search, an approach analogous to the one implemented by the PlasmidFinder web server (https://cge.cbs.dtu.dk/services/PlasmidFinder/) was applied. A script from the Center for Genomic Epidemiology core module (https://bitbucket.org/genomicepidemiology/cge_core_module) was used to run BLAST search and pre-process the hits resulting in one best hit per subject. The hits were then filtered based on the given cutoff values. At last, overlapping hits were removed. Plasmids with replicons having a corresponding pMLST scheme (IncA/C, IncF, IncHI1, IncHI2, IncI1 or IncN) were subjected to *in silico* pMLST analysis (13) using schemes and profiles from PubMLST (19) (https://pubmlst.org/plasmid/, 14 September 2018). The command line tool mlst, implemented by Seemann (https://github.com/tseemann/mlst, version 2.10), was applied using minimal identity of 85% and minimal coverage of 66% as recommended by Carattoli *et al.* (13). For the IncF plasmids, the sequence type was assigned according to the FAB formula (21). If the found allele hits were not exact (in terms of locus length and identity) or ambiguous (multiple exact hits) then the allele ID was not set. If more than one of the FIC/FII replicons had at least one exact allele match then the first part of the sequence type was set to '−−', i.e. ambiguous FIC/FII replicon hits; if none of these replicons had an exact allele hit then 'F-' was used. Next, Mash (15) (version 2.0) was applied to create sketches of the plasmid nucleotide sequences using parameters `-i -S 42 -p 20 -k 21 -s 1000`. The 2D embedding of the plasmid sequences was computed using UMAP (22) (version 0.2.5). First, pairwise distances between the sequences were computed from the created Mash sketches. Then, UMAP was applied to the distance matrix using parameters `n_neighbors=50, n_components=2, init='random', metric='precomputed'`. Unique pairs of similar plasmids were identified by computing pairwise distances with Mash with a distance cutoff of 0.00123693 which corresponds to have at least 950 of 1000 shared hashes. At last, a BLAST database was created using `makeblastdb` from the BLAST+ executables (14) (version 2.7.1+) called with the parameters `-input_type fasta -dbtype nucl`.

Overview of collected plasmids

In total, 13 789 plasmid records (2945 from INSDC and 10 844 from RefSeq) were retrieved from the NCBI nucleotide database. According to the date when the record was created, the number of plasmids increased drastically in the last years with more than 1000 unique sequences per year since 2015 (Figure 1). Moreover, the records collected since 2015 cover more than 60% of the dataset (9544 records). The sequence length of the obtained plasmid records ranged from 655 to 2 580 084 bp with a median of 52 830 bp. Furthermore, the created collection covered 1753 distinct species, 488 genera, 201 families, 98 orders, 42 classes and 22 phyla. The location coordinates could be obtained for 6171 records (44.8%). Using PlasmidFinder 5452 records (39.5%) could be characterized of which 2617 were subjected to *in silico* pMLST analysis.

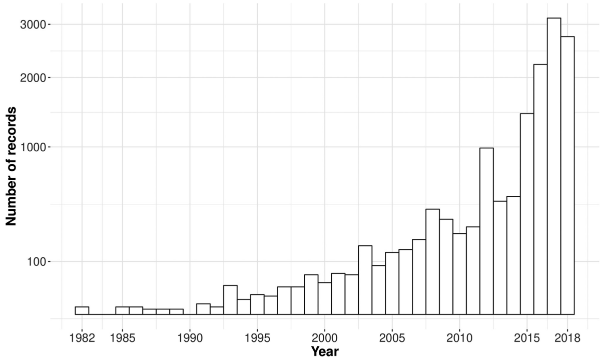


Figure 1. Number of plasmid records included into the collection grouped by the year of their creation. The *y*-axis scale is square root transformed.

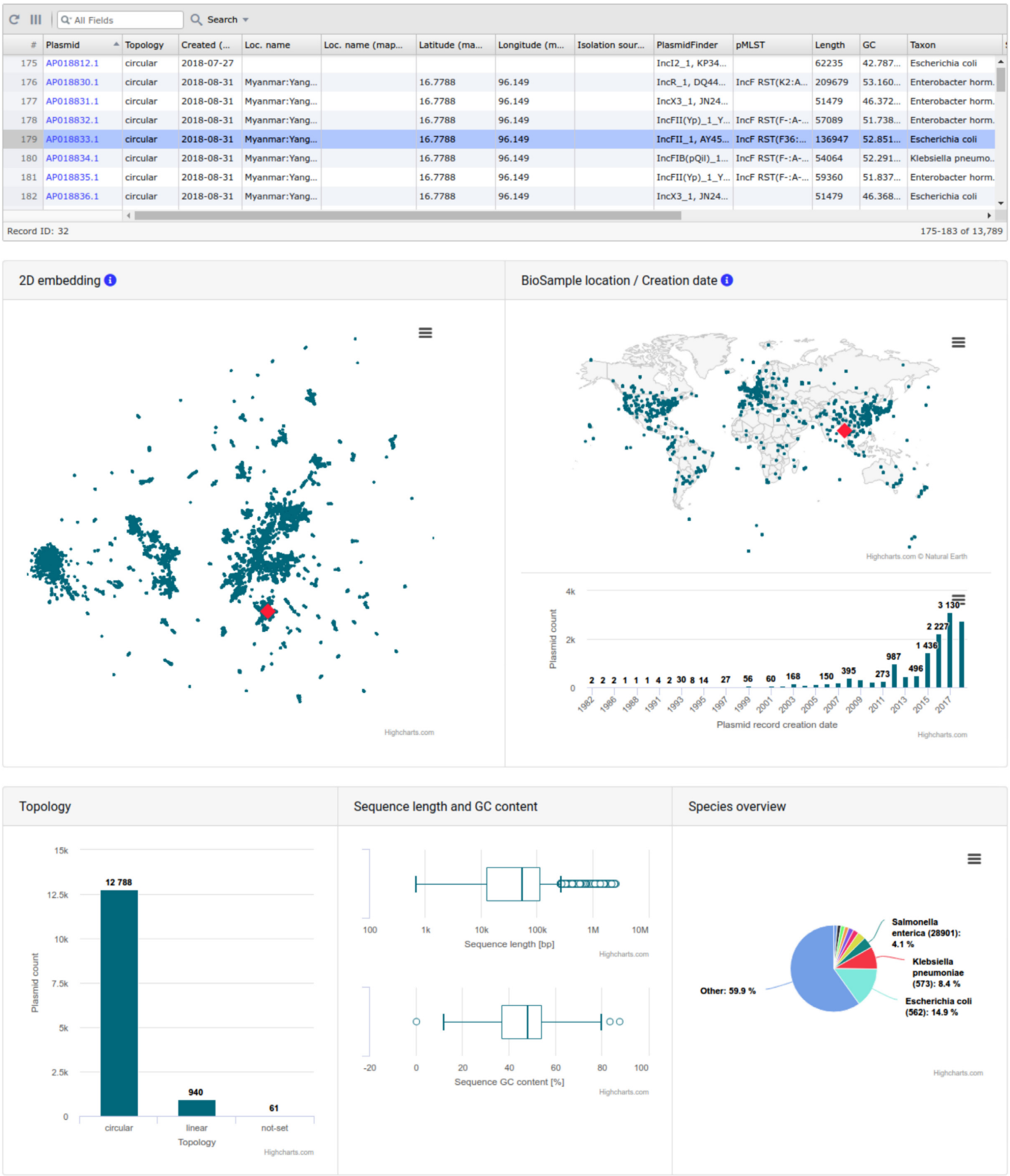
Resource implementation

The PLSDB was implemented as a document oriented resource using Django Python Web framework (https://djangoproject.com/) for the web server implementation. For user jobs, Celery (http://docs.celeryproject.org), a distributed task queue, is used together with Redis (https://redis.io/) as broker. The project was set up using CookieCutter (https://cookiecutter.readthedocs.io/) and Docker (https://www.docker.com/). Plots are drawn using the HighCharts library (https://www.highcharts.com/); the list of other used libraries can be found on the resource website. The resource update will be performed semi-automatically every 3 months together with the update of the used annotation databases. The web server code version, and the code version and date of data retrieval are provided for reference on the resource page.

DATABASE FUNCTIONALITY

Interactive overview of plasmids

A user-friendly and interactive view of the collected plasmid records is implemented (Figure 2). It includes a table showing the most relevant record information such as topology, record creation date, BioSample location and isolation source, PlasmidFinder and pMLST analysis results, nucleotide sequence length and GC content, and taxonomic information. Moreover, the 2D embedding of the records is shown together with a world map displaying records with available location information from the associated BioSample. At last, a summary of the shown records is provided including the number of records per year based on their creation date, sequence topology, the distribution of the sequence length and GC content, and the percentage of 10 most frequent species taxa. The taxonomic composition of all collected plasmid records is provided by an interactive Krona plot (23) showing the count and percentage of records for different taxa and ranks in the complete dataset and for each used resource (INSDC and RefSeq).



#	Plasmid	Topology	Created (...	Loc. name	Loc. name (map...	Latitude (ma...	Longitude (m...	Isolation sour...	PlasmidFinder	pMLST	Length	GC	Taxon
175	AP018812.1	circular	2018-07-27						IncI2_1, KP34...		62235	42.787...	Escherichia coli
176	AP018830.1	circular	2018-08-31	Myanmar:Yang...		16.7788	96.149		IncR_1, DQ44...	IncF RST(K2:A...	209679	53.160...	Enterobacter horm.
177	AP018831.1	circular	2018-08-31	Myanmar:Yang...		16.7788	96.149		IncX3_1, JN24...		51479	46.372...	Enterobacter horm.
178	AP018832.1	circular	2018-08-31	Myanmar:Yang...		16.7788	96.149		IncFII(Yp)_1_Y...	IncF RST(F-:A-...	57089	51.738...	Enterobacter horm.
179	AP018833.1	circular	2018-08-31	Myanmar:Yang...		16.7788	96.149		IncFII_1, AY45...	IncF RST(F36:...	136947	52.851...	Escherichia coli
180	AP018834.1	circular	2018-08-31	Myanmar:Yang...		16.7788	96.149		IncFIB(pQil)_1...	IncF RST(F-:A-...	54064	52.291...	Klebsiella pneumo..
181	AP018835.1	circular	2018-08-31	Myanmar:Yang...		16.7788	96.149		IncFII(Yp)_1_Y...	IncF RST(F-:A-...	59360	51.837...	Enterobacter horm.
182	AP018836.1	circular	2018-08-31	Myanmar:Yang...		16.7788	96.149		IncX3_1, JN24...		51479	46.368...	Escherichia coli

Record ID: 32

175-183 of 13,789

Figure 2. Interactive overview of collected plasmid records. Record AP018833.1 is selected in the table and highlighted (red diamond shaped symbol) in the embedding plot and on the world map.

The records can be filtered and searched through the table toolbar using any of the displayed table columns. The embedding, world map, and summary plots, except for the Krona plot, are then updated based on the filtering results. Each plasmid record has also a more detailed individual view which additionally includes plasmids associated with the same BioSample, plasmids being identical to the respective record (excluded from the dataset during the deduplication step) and similar plasmids (based on Mash distance), a table of hits to known resistance and virulence factors, and an interactive view of the sequence annotations provided through the NCBI sequence viewer (https://www.ncbi.nlm.nih.gov/projects/sviewer/).

Sequence search in plasmids

The PLSDB web server implements three options for sequence search: (i) short nucleotide sequences, e.g. genes, can be searched in the plasmid records using BLASTn (14). (ii) A potential plasmid represented by one or multiple nucleotide sequences (e.g. long or short reads, or contigs) can be searched in the resource by using Mash's distance estimation approach (15). Here, the sketches of the plasmid records are compared to the sketch of the uploaded sample to calculate their similarity. (iii) At last, the user can perform a containment analysis, also implemented by Mash (15). Here, the tool estimates the containment of each plasmid record in the uploaded nucleotide sequences by counting the number of shared hashes.

Application examples

In the following, we demonstrate how the PLSDB resource can be used in different scenarios for sequence data analysis.

Gene search. The first plasmid mediated bacterial resistance mechanism against colistin was reported by Liu *et al.* in 2016 describing the gene MCR-1 (3). The resistance factor was located on a plasmid found in an *Escherichia coli* strain extracted in the course of a surveillance project on antimicrobial resistance. The nucleotide sequence of MCR-1 (plasmid RefSeq accession KP347127.1, positions 22 413 to 24 038) was searched using BLASTn in the plasmid records with minimal identity and minimal query coverage per HSP set to 98% (Supplementary Table S1).

The search resulted in 253 hits. The plasmids included in the hits were mostly from *E. coli* (79.8%) and the remaining from other *Enterobacteriaceae* species; the corresponding records were included into the NCBI nucleotide database between 2015 and 2018. Most records were extracted from samples collected in China (31 of 58 records with location information) and most were labeled as collected from clinical patients (8 records of 51 records with isolation source information). In the latter case, the true number is likely to be higher as other labels (e.g. 'blood', 'urine', etc.) could also refer to clinical patient samples. The most frequently found replicons assigned by PlasmidFinder (13) were IncI2 (124 records), IncX4 (69 records) and IncHI2 (34 records). The retrieved plasmids could be used in a subsequent downscale analysis, e.g. by investigating the plasmids' genomic features in more detail.

Comparing plasmids. Li *et al.* (24) sequenced plasmids known to encode multi-drug resistance extracted from 12 bacterial strains (referred to as RB01 to RB12): 9 *E. coli*, 1 *Salmonella typhimurium*, 1 *Vibrio parahaemolyticus*, and 1 *Klebsiella pneumoniae*. In total, 21 plasmids could be assembled with 1–5 plasmids per sample for 11 of the 12 bacterial strains (sample R08, a *S. typhimurium*, was contaminated by chromosomal DNA). The nucleotide sequences of these plasmids were compared to the plasmid records stored in PLSDB using Mash (15) (command `dist`) with maximal *P*-value and distance thresholds set to 0.1 (Supplementary Table S2).

The taxonomy of the plasmid records from the best hit per query plasmid (hits were sorted by distance and number of shared hashes) matched the species taxon of the host bacteria in 15 of the 21 cases. Interestingly, two *E. coli* plasmids (from samples RB05 and RB06) had a perfect match (distance of 0, 1000 of 1000 shared hashes) to two distinct IncA/C2 plasmids extracted from *V. parahaemolyticus* (accessions MF627444.1 and MF627445.1). According to NCBI, these two *Vibrio* plasmids were found in cephalosporin-resistant *V. parahemolyticus* in retail shrimps in China. Both plasmids harbor the beta-lactamases $bla_{CTX-M-55}$ (ARO:3001917) and bla_{OXA-10} (ARO:3001405). However, the resistance factor CTX-M-15, also a beta-lactamase (ARO:3001878), present in samples RB05 and RB06, was not found in MF627444.1 or MF627445.1 (neither in the hits to known resistance factors nor in the feature names of the NCBI annotations) showing that there are differences in the gene content between the queries and the matched plasmids. These results demonstrate how the comparison analysis can help to identify potentially related plasmids found in other species.

Containment analysis. Schmidt *et al.* performed a study where they investigated the capability of MinION sequencing to identify pathogens in bacterial DNA enriched from urine of clinical patients (25). The raw MinION reads from this study were downloaded from the ENA web server (project accession PRJEB16761). From the included nine samples (CU4 - CU7, CU9, CU10, SU1, SU2 and S1D), only clinical urine (CU) samples were selected except for CU4 as its sequencing run was described as failed due to the poor quality of the used flow cells. The reads were extracted to FASTA files using Poretools (26) (version 0.6.0, `poretools fasta --type all reads.fast5 > reads.fasta`) and only the 'pass' reads were used for further analysis. For the five selected samples containment analysis was performed using Mash (15) (command `screen`) with maximal *P*-value set to 0.1 and minimal identity set to 0.99 (Supplementary Table S3).

From the five analyzed samples, hits were obtained only for CU6 and CU10. For CU6, plasmid records NZ_CP018990.1 and NZ_CP018964.1 were reported with 838 and 827 of 1000 shared hashes, respectively. Both plasmids were found in *E. coli*, were characterized as IncF plasmids and harbor multiple resistance factors including some of the genes found in CU6 by Schmidt *et al.*: *aadA5* and *dfrA17*. For CU10, one *E. coli* (NZ_CP011334.1) and five *K. pneumoniae* records (KY271405.1, KY271404.1,

Table 1. Comparison of pATLAS and PLSDB

Category	Sub-category	pATLAS[a]	PLSDB
Resource		RefSeq	RefSeq, **INSDC (DDBJ, EMBL-EBI, GenBank)**
Plasmid filtering		By specific words in FASTA header[b]	a query, genomic location and organism using edirect, by a regular expression on record description, completeness and taxonomy, de-duplication; removed putative chromosomal sequences
Number of plasmids		12 746	13 789
Plasmid overview	Presentation	**Distance-based network**, metadata table, summary plots	Metadata table, **embedding**, **world map**, summary plots, **Krona plot**
	Filtering	Sequence length, taxonomy, **annotations**	**Any column shown in metadata table**
Metadata	Sequence	Plasmid name, length, taxonomy	**Description/title** (incl. plasmid name), length, **GC content**, taxonomy, **topology**, **creation date**
	BioSample	✗	**Location**, **isolation source**
Annotation	ARG-Annot	✗	✓
	CARD	✓	✓
	ResFinder	✓	✓
	VFDB	✓	✓
	PlasmidFinder	✓	✓
	pMLST	✗	✓
Search	Local requirements	**Install and run provided pipeline**	**Download Mash sketches and BLAST DB, download tool binaries**
	Data upload	✗[c]	✓
Search strategy[d]	Mapping	✓(Bowtie2)	✗
	Distance estimation	✓(Mash)	✓(Mash)
	Containment	✓(Mash)	✓(Mash)
	Genes	✗	✓(BLASTn)

[a] pATLAS version 1.5.2 (last DB update from 20 July 2018), accessed on 1 August 2018.
[b] Derived from code review (https://github.com/tiagofilipe12/pATLAS/patlas/MASHix.py, commit `0f6dfa5`).
[c] Search results must be generated locally by the user using the pipeline provided by pATLAS. The results can be uploaded to the web server.
[d] For pATLAS, the information was derived from code review (https://github.com/tiagofilipe12/pATLASflow, commit `f3e9f2f`).
Bold text indicates differences between features; check mark indicates same/similar features and cross symbol a missing feature.

NZ_CP024500.1, NZ_CP024483.1 and NZ_CP024516.1) were obtained as hits. The *E. coli* plasmid was rather short with 2954 bp containing only four genomic annotations described as incomplete or frameshifted according to the NCBI nucleotide database. The five *K. pneumoniae* records were assigned to the same two replicons ('IncFIB(K)_1_Kpn3, JN233704' and 'IncFII(K)_1, CP000648') and were longer than 220 kbp except for KY271405.1 which was 133 069 bp. All of these five plasmids had hits to multiple resistance factors including genes identified in CU10 by Schmidt *et al.*: $bla_{CTX-M-15}$, bla_{OXA-1}, bla_{TEM}, *aac(6')Ib-cr*, *dfrA14*, *strB* and *qnrB* (more specifically *qnrB1*). These findings indicate a potential presence of plasmids bearing multiple antibiotic resistance factors in at least two of the analyzed samples and provide candidates for further analysis, e.g. to preform read alignment in order to determine whether the plasmids are fully covered, especially in the regions containing the resistance determinants.

COMPARISON TO EXISTING RESOURCES

The number of available resources providing a collection of known bacterial plasmids is limited.

The Addgene Repository is a database of plasmids generated by scientists and covering different organisms including bacteria (5). Though this is a highly extensive and valuable resource its purpose is not the compilation of naturally occurring bacterial plasmids but rather a platform for scientists to share plasmids used in the lab.

The PGD was created to include all fully sequenced plasmids (6). The records were collected from the NCBI database and included additionally to the bacterial plasmids also sequences from Archaea and Eukaryotes. But, this database is most likely not maintained anymore as its website is not accessible (http://www.genomics.ceh.ac.uk/plasmiddb/, accessed on 7 August 2018).

Orlek *et al.* (7) compiled a dataset of *Enterobacteriaceae* plasmids covering 2097 sequences in total and providing the protein sequences of translations in all six possible frames. However, this resource includes only data of a specific bacterial family and offers no web-based platform for data manipulation and analysis. The latter applies also to the dataset of 12 095 finished bacterial plasmids (accessed on 11 October 2018) created by Robertson and Nash for a software suit for processing plasmids from draft assemblies (8).

The pATLAS web server developed by Jesus, Gonçalves, Silva, Ramirez and Carriço (http://www.patlas.site, version 1.5.2, last DB update from 20 July 2018, accessed on 1 August 2018) includes bacterial plasmids extracted from NCBI RefSeq database (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plasmid/), annotated using ABRicate (https://github.com/tseemann/abricate) and compared using Mash (15). The plasmids are represented as a network where two plasmids are connected if their distance is below 0.1 and the associated *P*-value is below 0.05. The links in the network can be filtered and colored, and the nodes (i.e. plasmids) can be filtered by various parameters. Plasmids can be searched in high throughput sequencing data using Bowtie2 (27) (mapping based approach) or Mash (15) (distance estimation or containment analysis). In summary, pATLAS provides a comprehensive set of bacterial plasmids with an interactive network-based view and rich functionality. Compared to this resource, PLSDB additionally provides plasmid records from the INSDC resource which includes entries from DDBJ, EMBL-EBI and Genbank. As not all plasmids from INSDC are necessarily already included in RefSeq at the time of data retrieval, using both resources can provide a more complete set of records. Moreover, the meta-information in PLSDB includes further categories such as isolation location and source derived from the associated BioSamples. While pATLAS offers a mapping-based search which is not implemented in PLSDB, we offer the option to run a BLASTn search for short sequences, e.g. specific genetic markers such as resistance or virulence factors. Finally, in case of PLSDB, the user can upload the query sequences directly to the web-server. As the upload file size is limited, the required files can also be downloaded to run the search locally in case of having large datasets including many samples and/or sequences. A more detailed comparison of both resources can be found in Table 1.

CONCLUSION

The analysis of plasmids is essential for characterization of bacterial isolates and communities. Carrying different resistance and virulence factors, they also play a crucial role in dissemination of antibiotic resistance. We presented here PLSDB, an extensive resource of complete bacterial plasmids retrieved from the NCBI database. The implemented web server allows to browse the included plasmid records and to upload nucleotide sequences to be searched in the database using one of the three implemented options: search of short sequences such as genes, comparison of a plasmid sample to available plasmid records and containment analysis. The resource is freely accessible at https://ccb-microbe.cs.uni-saarland.de/plsdb.

CODE AND DATA AVAILABILITY

The code used to collect and process the data can be found at https://github.com/VGalata/plsdb. All relevant data files can be downloaded from the database website including plasmid metadata and annotations, mash sketches and BLAST database files. The resource can be accessed under the following URL: https://ccb-microbe.cs.uni-saarland.de/plsdb.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Saarland University. Funding for open access charge: Saarland University.

Conflict of interest statement. None declared.

REFERENCES

1. Couturier,M., Bex,F., Bergquist,P.L. and Maas,W.K. (1988) Identification and classification of bacterial plasmids. *Microbiol. Rev.*, **52**, 375–395.
2. Rozwandowicz,M., Brouwer,M.S.M., Fischer,J., Wagenaar,J.A., Gonzalez-Zorn,B., Guerra,B., Mevius,D.J. and Hordijk,J. (2018) Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J. Antimicrob. Chemother.*, **73**, 1121–1137.
3. Liu,Y.Y., Wang,Y., Walsh,T.R., Yi,L.X., Zhang,R., Spencer,J., Doi,Y., Tian,G., Dong,B., Huang,X. *et al.* (2016) Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect. Dis.*, **16**, 161–168.
4. Orlek,A., Phan,H., Sheppard,A.E., Doumith,M., Ellington,M., Peto,T., Crook,D., Walker,A.S., Woodford,N., Anjum,M.F. and Stoesser,N. (2017) Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid*, **91**, 42–52.
5. Kamens,J. (2015) The Addgene repository: an international nonprofit plasmid and data resource. *Nucleic Acids Res.*, **43**, D1152–D1157.
6. Mølbak,L., Tett,A., Ussery,D.W., Wall,K., Turner,S., Bailey,M. and Field,D. (2003) The plasmid genome database. *Microbiology (Reading, Engl.)*, **149**, 3043–3045.
7. Orlek,A., Phan,H., Sheppard,A.E., Doumith,M., Ellington,M., Peto,T., Crook,D., Walker,A.S., Woodford,N., Anjum,M.F. *et al.* (2017) A curated dataset of complete Enterobacteriaceae plasmids compiled from the NCBI nucleotide database. *Data Brief.*, **12**, 423–426.
8. Robertson,J. and Nash,J. H.E. (2018) MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genome*, **4**, doi:10.1099/mgen.0.000206.
9. Gupta,S.K., Padmanabhan,B.R., Diene,S.M., Lopez-Rojas,R., Kempf,M., Landraud,L. and Rolain,J.M. (2014) ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.*, **58**, 212–220.
10. Jia,B., Raphenya,A.R., Alcock,B., Waglechner,N., Guo,P., Tsang,K.K., Lago,B.A., Dave,B.M., Pereira,S., Sharma,A.N. *et al.* (2017) CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **45**, D566–D573.
11. Zankari,E., Hasman,H., Cosentino,S., Vestergaard,M., Rasmussen,S., Lund,O., Aarestrup,F.M. and Larsen,M.V. (2012) Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.*, **67**, 2640–2644.
12. Chen,L., Yang,J., Yu,J., Yao,Z., Sun,L., Shen,Y. and Jin,Q. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, **33**, D325–D328.
13. Carattoli,A., Zankari,E., Garcia-Fernandez,A., Voldby Larsen,M., Lund,O., Villa,L., Møller Aarestrup,F. and Hasman,H. (2014) In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.*, **58**, 3895–3903.
14. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
15. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
16. Broder,A. (1998) On the resemblance and containment of documents. In: Carpentieri,B (ed). *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*. IEEE Computer Society, Los Alamitos, pp. 21–29.
17. Kans,J. (2013) *Entrez Direct: E-utilities on the UNIX Command Line*. National Center for Biotechnology Information (US), Bethesda, MD.
18. Jolley,K.A., Bliss,C.M., Bennett,J.S., Bratcher,H.B., Brehony,C., Colles,F.M., Wimalarathna,H., Harrison,O.B., Sheppard,S.K., Cody,A.J. *et al.* (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology (Reading, Engl.)*, **158**, 1005–1015.

19. Jolley,K.A. and Maiden,M.C. (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, **11**, 595.
20. Yutin,N., Puigbo,P., Koonin,E.V. and Wolf,Y.I. (2012) Phylogenomics of prokaryotic ribosomal proteins. *PLoS One*, **7**, e36972.
21. Villa,L., Garcia-Fernandez,A., Fortini,D. and Carattoli,A. (2010) Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. *J. Antimicrob. Chemother.*, **65**, 2518–2529.
22. McInnes,L., Healy,J., Saul,N. and Großberger,L. (2018) UMAP: uniform manifold approximation and projection. *J. Open Source Softw.*, **3**, 861.
23. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
24. Li,R., Xie,M., Dong,N., Lin,D., Yang,X., Wong,M. H.Y., Chan,E.W. and Chen,S. (2018) Efficient generation of complete sequences of MDR-encoding plasmids by rapid assembly of MinION barcoding sequencing data. *Gigascience*, **7**, 1–9.
25. Schmidt,K., Mwaigwisya,S., Crossman,L.C., Doumith,M., Munroe,D., Pires,C., Khan,A.M., Woodford,N., Saunders,N.J., Wain,J. *et al.* (2017) Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J. Antimicrob. Chemother.*, **72**, 104–114.
26. Loman,N.J. and Quinlan,A.R. (2014) Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, **30**, 3399–3401.
27. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

4
Discussion and conclusions

WGS is increasingly applied in microbiology research to study microbial communities and bacterial isolates. In particular, the study of antibiotic resistance has gained in importance due to the emergence of highly resistant bacterial pathogens. The growing amount of data requires the development of reference databases and new tools to facilitate the analysis.

This thesis covered the analysis of a collection of bacterial isolates and the implementation of computational resources to process microbial genomic data with a strong focus on bacterial pathogens and antibiotic resistance.

GEAR-base project A large-scale dataset of around eleven thousand clinical bacterial isolates collected at different locations over the world and over a time span of 30 years allowed for an analysis of antibiotic resistance data and genomic features from different bacterial species known to be relevant human pathogens. First, WGS-based tools for taxonomic classification were evaluated on a subset of samples of the isolate collection, for which MS-based taxonomic information was available. The Kraken tool, which was among the best-performing tools, was used to classify all the isolates of the GEAR-base project, providing a consistent WGS-based taxonomic characterization. Then, various analyses of the resistance profiles, meta-information and genomic data were performed, focusing on different bacterial species. An online resource, GEAR-base, was also implemented, providing the collected and generated data through three main modules: the culture-based module, the pan-genome module, and the analysis module. The culture-based module included an overview of the antibiotic resistance and sample metadata. It provides summaries of median MIC values, resistance percentages, and the number of resistant and susceptible samples with respect to different taxa and taxonomy ranks, collection location and collection year. The pan-genome module contains the newly constructed pan-genomes, including all identified gene clusters (centroids). The centroids of a pan-genome can be filtered by their frequency in the sample cohort and function, and resistance-related information is shown for each centroid, if it was collected originally. The analysis module implements different options for a search of genes or genomes uploaded by the user in the data included in the resource.

No apparently new resistance factors could be discovered in the dataset which indicates that these are either not contained in the data or that the chosen methods were not able to detect them, e.g. because of their low frequency in the samples. Nevertheless, GEAR-base constitutes a valuable data resource as it provides an extensive collection covering relevant human pathogens and antibiotics, and further studies can be performed to potentially complement the hitherto identified resistance factors. The herein described centroid-based approach considers only the presence of coding sequences in bacterial genomes. However, mutations and non-coding genetic regions can also contribute to the development of a resistance phenotype. Using *k*-mers (subsequences of length *k*) would allow for finding, besides the presence of specific genes, mutations and other small structural changes in coding and non-coding sequences. Recently, multiple tools have been published for performing genome-wide association analyses using *k*-mers for large bacterial datasets [236–238]. Additionally, not only taking into account the presence of genomic features but also their frequency in the genomes would enable the identification of resistance factors whose copy number can be correlated with the resistance phenotype. The insights gained from the available data and new analyses can be used to determine sets of genomic features for predicting antibiotic resistance. This kind of analysis was done by Davis *et al.* , who trained *k*-mer-based classifiers using bacterial genomes from the PATRIC database [239]. The *in silico* identified genomic features can then serve as a basis for the development of resistance tests to be applied in clinic and for antibiotic surveillance purposes.

Since new resistance determinants can emerge and spread, it is crucial to keep a database storing the related data up-to-date. It would therefore be of great benefit to extend the GEAR-base isolate collection by new clinical samples provided by industrial and academic partners, and healthcare facilities. As metagenomics and long-read sequencing are increasingly applied to study antibiotic resistance and bacterial infections, including such data into the resource requires the selection and implementation of suitable analysis approaches. In both cases, dedicated tools are required for sequencing data processing, taxonomic profiling, *de novo* assembly and annotation, and further downstream analysis.

Metagenomic binning One of the main tasks in the processing of metagenomic data is binning of the given genomic sequences into clusters, which are then used in further downstream analyses. The web server BusyBee Web provides an online tool for metagenomic binning, which does not require any further data besides the nucleotide sequences of a metagenomic sample. The implemented approach is based on sequence derived features (i.e. five-mer frequencies), a non-linear embedding of the five-mer profiles into a two-dimensional space, and a bootstrap supervised binning step, where clusters are learned from a pre-defined subset of the input sequences and the remaining sequences are assigned to these clusters to create the final bins. Additionally, the bins are evaluated to estimate their completeness and purity, and the results can also include taxonomic annotation and detection of known resistance genes from the Resfams database.

One shortcoming of the current implementation of the web server is that it allowed to process only one metagenomic sample, without offering the possibility to compare the results of multiple jobs, or to upload a multi-sample dataset. But, comparing sample groups or data from different time points is a frequent objective in metagenomic studies. The problem which arises when uploading multiple samples together would be a significant increase to the running time for the embedding step. A potential solution would be to perform the embedding on a subset of the input sequences, and to project the remaining sequences into the computed space. While the currently used implementation of the embedding BH-SNE approach [240] does not provide methods for accomplishing this task, a latterly introduced UMAP approach [241] offers this functionality. The results of the multi-sample analysis can be further processed to highlight the distribution of the sample-sequences in the embedding plot and to compare the found organisms between the samples or sample groups.

The available sequence and bin annotation options can be further extended to provide a more detailed characterization of the analyzed sample. These could include other resistance gene databases, such as the Comprehensive Antibiotic Resistance Database (CARD) [242], virulence factor databases, such as the Virulence Factor Database (VFDB) [243], and plasmid databases, such as the herein presented PLSDB resource [5].

Bacterial plasmids Bacterial plasmids are one of the key factors in antibiotic resistance emergence and spread. Currently, there are various tools for classifying genomic sequences as originating from a chromosome or a plasmid. Four plasmid prediction tools were reviewed, covering three different approaches (assembly-based, marker based, and *k*-mer based) and applied to a sample subset from the GEAR-base isolate collection. The results demonstrated high heterogeneity between the tested tools and taxon-dependent variation.

As there is no gold-standard approach to recovering plasmid sequences from WGS data, it is desirable to at least be able to detect known plasmids. This would require a collection of known bacterial plasmids which could also be used for the identification and characterization of "pure" plasmid samples. To this end, we implemented the PLSDB resource, which contains complete bacterial plasmids submitted to NCBI. In addition to the plasmid records provided, this resource also includes further metadata for the related biological sample and assembly records, and annotation information such as plasmid typing, resistance and virulence genes. The PLSDB resource offers multiple analysis options for genomic data uploaded by a user: detection of known plasmids in WGS data of bacterial isolates or metagenomes, comparison of plasmids to the ones stored in the resource, and listing plasmids containing given genomic features (e.g. specific genes).

While PLSDB provides a view of the aggregated meta-information of a set of plasmids, there is currently no option for comparing their genetic content. The plasmid typing data (replicon typing and pMLST [234]) and the estimated genetic dissimilarity (using a *k*-mer and MinHash-based tool Mash [244]) constitute only an approximation of true relationships between the plasmids. A more comprehensive approach is to represent the shared genetic content of the plasmids through a network [245]. This can be accomplished by using a variation graph which encodes a set of sequences (e.g. genomes), reflecting their genetic differences (variations) [246]. The plasmids stored in PLSDB could be used to construct such variation graphs for the plasmid groups of interest, to study their relationships, and to be used as references for whole-genome sequencing data.

In summary, in this thesis, we have presented multiple online resources which provide valuable data collections and/or tools for the analysis of microbial or bacterial genomic data, with a focus on antibiotic resistance. Given the increased use of sequencing data in microbial research, the herein presented body of work is expected to benefit the community through the ease-of-use and breadth of these data collections and tools.

Bibliography

[1] V. Galata, C. Backes, C. C. Laczny, G. Hemmrich-Stanisak, H. Li, L. Smoot, A. E. Posch, S. Schmolke, M. Bischoff, L. von Muller, A. Plum, A. Franke, and A. Keller. Comparing genome versus proteome-based identification of clinical bacterial isolates. *Brief. Bioinformatics*, 19(3):495–505, May 2018.

[2] V. Galata, C. C. Laczny, C. Backes, G. Hemmrich-Stanisak, S. Schmolke, A. Franke, E. Meese, M. Herrmann, L. von Muller, A. Plum, R. Muller, C. Stahler, A. E. Posch, and A. Keller. Integrating Culture-based Antibiotic Resistance Profiles with Whole-genome Sequencing Data for 11,087 Clinical Isolates. *Genomics Proteomics Bioinformatics*, 17(2):169–182, Apr 2019.

[3] C. C. Laczny, C. Kiefer, V. Galata, T. Fehlmann, C. Backes, and A. Keller. BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res.*, 45(W1):W171–W179, Jul 2017.

[4] C. C. Laczny, V. Galata, A. Plum, A. E. Posch, and A. Keller. Assessing the heterogeneity of in silico plasmid predictions based on whole-genome-sequenced clinical isolates. *Brief. Bioinformatics*, 20(3):857–865, May 2019.

[5] V. Galata, T. Fehlmann, C. Backes, and A. Keller. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.*, 47 (D1):D195–D202, Jan 2019.

[6] John A Fuerst. Beyond Prokaryotes and Eukaryotes : Planctomycetes and Cell Organization. *Nature Education*, 3(9):44, 2010.

[7] C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.*, 87(12):4576–4579, Jun 1990.

[8] R. M. Pelczar and M. J. Pelczar. Encyclopædia Britannica: Microbiology, Jan 2018. URL `https://www.britannica.com/science/microbiology`. Accessed: 12.08.2019.

[9] J. Kallmeyer, R. Pockalny, R. R. Adhikari, D. C. Smith, and S. D'Hondt. Global distribution of microbial abundance and

biomass in subseafloor sediment. *Proc. Natl. Acad. Sci. U.S.A.*, 109(40):16213–16216, Oct 2012.

[10] C. W. Sullivan and A. C. Palmisano. Sea Ice Microbial Communities: Distribution, Abundance, and Diversity of Ice Bacteria in McMurdo Sound, Antarctica, in 1980. *Appl. Environ. Microbiol.*, 47(4):788–795, Apr 1984.

[11] P. H. Rampelotto. Extremophiles and extreme environments. *Life (Basel)*, 3(3):482–485, Aug 2013.

[12] L. J. Rothschild and R. L. Mancinelli. Life in extreme environments. *Nature*, 409(6823):1092–1101, Feb 2001.

[13] A. Moya, R. Gil, and A. Latorre. The evolutionary history of symbiotic associations among bacteria and their animal hosts: a model. *Clin. Microbiol. Infect.*, 15 Suppl 1:11–13, Jan 2009.

[14] A. L. Byrd, Y. Belkaid, and J. A. Segre. The human skin microbiome. *Nat. Rev. Microbiol.*, 16(3):143–155, Mar 2018.

[15] L. Wen and A. Duffy. Factors Influencing the Gut Microbiota, Inflammation, and Type 2 Diabetes. *J. Nutr.*, 147(7):1468S–1475S, Jul 2017.

[16] L. V. Hooper and J. I. Gordon. Commensal host-bacterial relationships in the gut. *Science*, 292(5519):1115–1118, May 2001.

[17] J. R. Marchesi and J. Ravel. The vocabulary of microbiome research: a proposal. *Microbiome*, 3:31, 2015.

[18] R. M. Stubbendieck, C. Vargas-Bautista, and P. D. Straight. Bacterial Communities: Interactions to Scale. *Front Microbiol*, 7: 1234, 2016.

[19] D. Gevers, R. Knight, J. F. Petrosino, K. Huang, A. L. McGuire, B. W. Birren, K. E. Nelson, O. White, B. A. Methe, and C. Huttenhower. The Human Microbiome Project: a community resource for the healthy human microbiome. *PLoS Biol.*, 10(8):e1001377, 2012.

[20] R. Sender, S. Fuchs, and R. Milo. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.*, 14(8):e1002533, Aug 2016.

[21] C. N. Spaulding, R. D. Klein, H. L. Schreiber, J. W. Janetka, and S. J. Hultgren. Precision antimicrobial therapeutics: the path of least resistance? *NPJ Biofilms Microbiomes*, 4:4, 2018.

[22] A. Schulfer and M. J. Blaser. Risks of Antibiotic Exposures Early in Life on the Developing Microbiome. *PLoS Pathog.*, 11 (7):e1004903, Jul 2015.

[23] J. R. Porter and A. van Leeuwenhoek. Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. *Bacteriol Rev*, 40 (2):260–269, Jun 1976.

[24] S. M. Blevins, M. S. Bronze, and R. Koch. Robert Koch and the 'golden age' of bacteriology. *Int. J. Infect. Dis.*, 14(9):e744–751, Sep 2010.

[25] D. Medini, D. Serruto, J. Parkhill, D. A. Relman, C. Donati, R. Moxon, S. Falkow, and R. Rappuoli. Microbiology in the post-genomic era. *Nat. Rev. Microbiol.*, 6(6):419–430, Jun 2008.

[26] N. J. Palleroni. Prokaryote taxonomy of the 20th century and the impact of studies on the genus Pseudomonas: a personal view. *Microbiology (Reading, Engl.)*, 149(Pt 1):1–7, Jan 2003.

[27] Wayne, L. G. and Brenner, D. J. and Colwell, R. R. and Grimont, P. A. D. and Kandler, o. and Krichevsky, M. I. and Moore, L. H. and Moore, W. E. C. and Murray, R. G. E. and Stackebrandt, E. and Starr, M. P. an Truper, H. G. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic Bacteriology*, 37(4):463–464, Oct 1987.

[28] E. Stackebrandt, W. Frederiksen, G. M. Garrity, P. A. Grimont, P. Kampfer, M. C. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.*, 52(Pt 3):1043–1047, May 2002.

[29] R. Rossello-Mora and R. Amann. The species concept for prokaryotes. *FEMS Microbiol. Rev.*, 25(1):39–67, Jan 2001.

[30] B. J. Tindall, R. Rossello-Mora, H. J. Busse, W. Ludwig, and P. Kampfer. Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol.*, 60(Pt 1): 249–266, Jan 2010.

[31] G. M. Garrity. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? *J. Clin. Microbiol.*, 54(8): 1956–1963, Aug 2016.

[32] T. Coenye, D. Gevers, Y. Van de Peer, P. Vandamme, and J. Swings. Towards a prokaryotic genomic taxonomy. *FEMS Microbiol. Rev.*, 29(2):147–167, Apr 2005.

[33] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*, 9(1): 5114, Nov 2018.

[34] D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P. A. Chaumeil, and P. Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, 36(10):996–1004, Nov 2018.

[35] M. T. Madigan, J. M. Martinko, and J. Parker. *Brock Biology of Microorganisms*. Prentice-Hall, 8th edition, 1997. ISBN 0-13-571225-4.

[36] I. A. Khan. Plague: the dreadful visitation occupying the human mind for centuries. *Trans. R. Soc. Trop. Med. Hyg.*, 98(5):270–277, May 2004.

[37] T. D. Brock. *Robert Koch: a life in medicine and bacteriology*. Zondervan, Washington, DC: American Society of Microbiology Press, 1999.

[38] B. Sobolewska, M. Buhl, J. Liese, and F. Ziemssen. Slit lamps and lenses: a potential source of nosocomial infections? *Eye (Lond)*, 32(6):1021–1027, Jun 2018.

[39] H. C. Steel, R. Cockeran, R. Anderson, and C. Feldman. Overview of community-acquired pneumonia and the role of inflammatory mechanisms in the immunopathogenesis of severe pneumococcal disease. *Mediators Inflamm.*, 2013:490346, 2013.

[40] W. O. C. M. Cookson, M. J. Cox, and M. F. Moffatt. New opportunities for managing acute and chronic lung infections. *Nat. Rev. Microbiol.*, 16(2):111–120, Feb 2018.

[41] E. Martens and A. L. Demain. The antibiotic resistance crisis, with a focus on the United States. *J. Antibiot.*, 70(5):520–526, May 2017.

[42] M. Davenport, K. E. Mach, L. M. D. Shortliffe, N. Banaei, T. H. Wang, and J. C. Liao. New and developing diagnostic technologies for urinary tract infections. *Nat Rev Urol*, 14(5):296–310, May 2017.

[43] C. R. Arciola, D. Campoccia, and L. Montanaro. Implant infections: adhesion, biofilm formation and immune evasion. *Nat. Rev. Microbiol.*, 16(7):397–409, Jul 2018.

[44] European Centre for Disease Prevention and Control (ECDC). Vaccine schedules in all countries of the European Union, 2019. URL `https://vaccine-schedule.ecdc.europa.eu/`. Accessed: 12.06.2019.

[45] Centers for Disease Control and Prevention (CDC). Immunization schedules, 2019. URL `https://www.cdc.gov/vaccines/schedules/index.html`. Accessed: 12.06.2019.

[46] F. Baquero, A. P. Tedim, and T. M. Coque. Antibiotic resistance shaping multi-level population biology of bacteria. *Front Microbiol*, 4:15, 2013.

[47] B. L. Ligon, A. Fleming, H. W. Florey, and E. B. Chain. Penicillin: its discovery and early development. *Semin Pediatr Infect Dis*, 15(1):52–57, Jan 2004.

[48] I. B. Seiple, Z. Zhang, P. Jakubec, A. Langlois-Mercier, P. M. Wright, D. T. Hog, K. Yabu, S. R. Allu, T. Fukuzaki, P. N. Carlsen, Y. Kitamura, X. Zhou, M. L. Condakes, F. T. Szczypiński, W. D. Green, and A. G. Myers. A platform for the discovery of new macrolide antibiotics. *Nature*, 533(7603):338–345, May 2016.

[49] S. Chakraborty and K. Y. Rhee. Tuberculosis Drug Development: History and Evolution of the Mechanism-Based Paradigm. *Cold Spring Harb Perspect Med*, 5(8):a021147, Apr 2015.

[50] H. Irschik, K. Gerth, G. Hofle, W. Kohl, and H. Reichenbach. The myxopyronins, new inhibitors of bacterial RNA synthesis from Myxococcus fulvus (Myxobacterales). *J. Antibiot.*, 36(12): 1651–1658, Dec 1983.

[51] W. Snaeder. *Drug Discovery: A History*. John Wiley & Sons Inc, West Sussex, England, 2005. ISBN 0471899798.

[52] D. Hendlin, E. O. Stapley, M. Jackson, H. Wallick, A. K. Miller, F. J. Wolf, T. W. Miller, L. Chaiet, F. M. Kahan, E. L. Foltz, H. B. Woodruff, J. M. Mata, S. Hernandez, and S. Mochales. Phosphonomycin, a new antibiotic produced by strains of streptomyces. *Science*, 166(3901):122–123, Oct 1969.

[53] W. O. Godtfredsen, S. Jahnsen, H. Lorck, K. Roholt, and L. Tybring. Fusidic acid: a new antibiotic. *Nature*, 193:987, Mar 1962.

[54] S. Gurusiddaiah and S. O. Graham. Some chemical and physical characteristics of pantomycin, and antiobiotic isolated from Streptomyces hygroscopicus. *Antimicrob. Agents Chemother.*, 17 (6):980–987, Jun 1980.

[55] L. E. Cooper, B. Li, and W. A. van der Donk. *Comprehensive Natural Products II*, chapter "5.08- Biosynthesis and Mode of Action of Lantibiotics", pages 217 – 256. Elsevier, Oxford, 2010. ISBN 978-0-08-045382-8. doi: https://doi.org/10.1016/B978-008045382-8.00116-7.

[56] D. J. Mason, A. Dietz, and C. DeBoer. Lincomycin, a new antibiotic. i. Discovery and biological properties. *Antimicrobial Agents and Chemotherapy*, pages 554 – 559, 1962.

[57] F. Parenti. Structure and mechanism of action of teicoplanin. *J. Hosp. Infect.*, 7 Suppl A:79–83, Mar 1986.

[58] G. M. Eliopoulos, S. Willey, E. Reiszner, P. G. Spitzer, G. Caputo, and R. C. Moellering. In vitro and in vivo activity of LY 146032, a new cyclic lipopeptide antibiotic. *Antimicrob. Agents Chemother.*, 30(4):532–535, Oct 1986.

[59] A. T. Fuller, G. Mellows, M. Woolford, G. T. Banks, K. D. Barrow, and E. B. Chain. Pseudomonic acid: an antibiotic produced by Pseudomonas fluorescens. *Nature*, 234(5329):416–417, Dec 1971.

[60] M. C. Dodd, W. B. Stillman, M. Roys, and C. Crosby. The in vitro bacteriostatic action of some simple furan derivatives. *Journal of Pharmacology and Experimental Therapeutics*, 82(1):11–18, 1944. ISSN 0022-3565.

[61] K. Maeda, T. Osato, and H. Umezawa. A new antibiotic, azomycin. *J. Antibiot.*, 6(4):182, Dec 1953.

[62] B. Bozdogan and P. C. Appelbaum. Oxazolidinones: activity, mode of action, and mechanism of resistance. *Int. J. Antimicrob. Agents*, 23(2):113–119, Feb 2004.

[63] J. Wang, S. M. Soisson, K. Young, W. Shoop, S. Kodali, A. Galgoci, R. Painter, G. Parthasarathy, Y. S. Tang, R. Cummings, S. Ha, K. Dorso, M. Motyl, H. Jayasuriya, J. Ondeyka, K. Herath, C. Zhang, L. Hernandez, J. Allocco, A. Basilio, J. R. Tormo, O. Genilloud, F. Vicente, F. Pelaez, L. Colwell, S. H. Lee, B. Michael, T. Felcetto, C. Gill, L. L. Silver, J. D. Hermes, K. Bartizal, J. Barrett, D. Schmatz, J. W. Becker, D. Cully, and S. B. Singh. Platensimycin is a selective FabF inhibitor with potent antibiotic properties. *Nature*, 441(7091):358–361, May 2006.

[64] R. Novak and D. M. Shlaes. The pleuromutilin antibiotics: a new class for human use. *Curr Opin Investig Drugs*, 11(2): 182–191, Feb 2010.

[65] G. C. AINSWORTH, A. M. BROWN, and G. BROWNLEE. Aerosporin, an antibiotic produced by Bacillus aerosporus Greer. *Nature*, 159(4060):263, Aug 1947.

[66] H. Takahashi, I. Hayakawa, and T. Akimoto. [The history of the development and changes of quinolone antibacterial agents]. *Yakushigaku Zasshi*, 38(2):161–179, 2003.

[67] J. CHARNEY, W. P. FISHER, C. CURRAN, R. A. MACHLOWITZ, and A. A. TYTELL. Streptogramin, a new antibiotic. *Antibiot Chemother (Northfield)*, 3(12):1283–1286, Dec 1953.

[68] C. DEBOER, A. DIETZ, G. M. SAVAGE, and W. S. SILVER. Streptolydigin, a new antimicrobial antibiotic. I. Biologic studies of streptolydigin. *Antibiot Annu*, 3:886–892, 1955.

[69] G. M. Eliopoulos and P. Huovinen. Resistance to trimethoprim-sulfamethoxazole. *Clin. Infect. Dis.*, 32(11):1608–1614, Jun 2001. ISSN 1058-4838. doi: 10.1086/320532.

[70] L. L. Ling, T. Schneider, A. J. Peoples, A. L. Spoering, I. Engels, B. P. Conlon, A. Mueller, T. F. Schaberle, D. E. Hughes, S. Epstein, M. Jones, L. Lazarides, V. A. Steadman, D. R. Cohen, C. R. Felix, K. A. Fetterman, W. P. Millett, A. G. Nitti, A. M. Zullo, C. Chen, and K. Lewis. A new antibiotic kills pathogens without detectable resistance. *Nature*, 517(7535):455–459, Jan 2015.

[71] R. G. COOPER and M. WALD. Successful treatment of proteus septicaemia with a new drug trimethoprim. *Med. J. Aust.*, 2: 93–96, Jul 1964.

[72] B. M. Hover, S. H. Kim, M. Katz, Z. Charlop-Powers, J. G. Owen, M. A. Ternei, J. Maniko, A. B. Estrela, H. Molina, S. Park, D. S. Perlin, and S. F. Brady. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nat Microbiol*, 3(4):415–422, Apr 2018.

[73] G. R. Donowitz and G. L. Mandell. Beta-Lactam antibiotics (1). *N. Engl. J. Med.*, 318(7):419–426, Feb 1988.

[74] A. C. Chien, N. S. Hill, and P. A. Levin. Cell size control in bacteria. *Curr. Biol.*, 22(9):R340–349, May 2012.

[75] K. D. Young. The selective value of bacterial shape. *Microbiol. Mol. Biol. Rev.*, 70(3):660–703, Sep 2006.

[76] E. J. Baron. *Medical Microbiology*. University of Texas Medical Branch at Galveston, 4th edition, 1996.

[77] H. Wu, C. Moser, H. Z. Wang, N. Høiby, and Z. J. Song. Strategies for combating bacterial biofilm infections. *Int J Oral Sci*, 7 (1):1–7, Mar 2015.

[78] I-Hsiu Huang, Prabhat Dwivedi, and Hung Ton-That. *Bacterial Pili and Fimbriae*. American Cancer Society, 2010. ISBN 9780470015902. doi: 10.1002/9780470015902.a0000304.pub2.

[79] C. A. Kerfeld, C. Aussignargues, J. Zarzycki, F. Cai, and M. Sutter. Bacterial microcompartments. *Nat. Rev. Microbiol.*, 16(5): 277–290, May 2018.

[80] M. Choudhary, C. Mackenzie, K. S. Nereng, E. Sodergren, G. M. Weinstock, and S. Kaplan. Multiple chromosomes in bacteria: structure and function of chromosome II of Rhodobacter sphaeroides 2.4.1T. *J. Bacteriol.*, 176(24):7694–7702, Dec 1994.

[81] M. Y. Galperin. Linear chromosomes in bacteria: no straight edge advantage? *Environ. Microbiol.*, 9(6):1357–1362, Jun 2007.

[82] M. Thanbichler, P. H. Viollier, and L. Shapiro. The structure and function of the bacterial chromosome. *Curr. Opin. Genet. Dev.*, 15(2):153–162, Apr 2005.

[83] S. C. Dillon and C. J. Dorman. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat. Rev. Microbiol.*, 8(3):185–195, Mar 2010.

[84] L. Postow, C. D. Hardy, J. Arsuaga, and N. R. Cozzarelli. Topological domain structure of the Escherichia coli chromosome. *Genes Dev.*, 18(14):1766–1779, Jul 2004.

[85] D. J. Sherratt. Bacterial chromosome dynamics. *Science*, 301 (5634):780–785, Aug 2003.

[86] M. Valens, S. Penaud, M. Rossignol, F. Cornet, and F. Boccard. Macrodomain organization of the Escherichia coli chromosome. *EMBO J.*, 23(21):4330–4341, Oct 2004.

[87] M. T. Cabeen and C. Jacobs-Wagner. Bacterial cell shape. *Nat. Rev. Microbiol.*, 3(8):601–610, Aug 2005.

[88] E. Sauvage, F. Kerff, M. Terrak, J. A. Ayala, and P. Charlier. The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. *FEMS Microbiol. Rev.*, 32(2):234–258, Mar 2008.

[89] L. L. Silver. Multi-targeting by monotherapeutic antibacterials. *Nat Rev Drug Discov*, 6(1):41–55, Jan 2007.

[90] P. E. Reynolds. Structure, biochemistry and mechanism of action of glycopeptide antibiotics. *Eur. J. Clin. Microbiol. Infect. Dis.*, 8(11):943–950, Nov 1989.

[91] L. Friedman, J. D. Alder, and J. A. Silverman. Genetic changes that correlate with reduced susceptibility to daptomycin in Staphylococcus aureus. *Antimicrob. Agents Chemother.*, 50(6): 2137–2145, Jun 2006.

[92] C. C. HsuChen and D. S. Feingold. The mechanism of polymyxin B action and selectivity toward biologic membranes. *Biochemistry*, 12(11):2105–2111, May 1973.

[93] G. Charvin, D. Bensimon, and V. Croquette. Single-molecule study of DNA unlinking by eukaryotic and prokaryotic type-II topoisomerases. *Proc. Natl. Acad. Sci. U.S.A.*, 100(17):9820–9825, Aug 2003.

[94] J. M. Dewar and J. C. Walter. Mechanisms of DNA replication termination. *Nat. Rev. Mol. Cell Biol.*, 18(8):507–516, Aug 2017.

[95] M. Boolchandani, A. W. D'Souza, and G. Dantas. Sequencing-based methods and resources to study antimicrobial resistance. *Nat. Rev. Genet.*, 20(6):356–370, Jun 2019.

[96] L. Postow, N. J. Crisona, B. J. Peter, C. D. Hardy, and N. R. Cozzarelli. Topological challenges to DNA replication: conformations at the fork. *Proc. Natl. Acad. Sci. U.S.A.*, 98(15): 8219–8226, Jul 2001.

[97] S. Hawser, S. Lociuro, and K. Islam. Dihydrofolate reductase inhibitors as antibacterial agents. *Biochem. Pharmacol.*, 71(7): 941–948, Mar 2006.

[98] J. Vila. *Frontiers in Antimicrobial Resistance*, chapter "Fluoroquinolone Resistance", pages 41–52. American Society of Microbiology, 2005. doi: 10.1128/9781555817572.ch4.

[99] T. A. Steitz. A structural understanding of the dynamic ribosome machine. *Nat. Rev. Mol. Cell Biol.*, 9(3):242–253, Mar 2008.

[100] L. S. Redgrave, S. B. Sutton, M. A. Webber, and L. J. Piddock. Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. *Trends Microbiol.*, 22(8):438–445, Aug 2014.

[101] K. J. Aldred, R. J. Kerns, and N. Osheroff. Mechanism of quinolone action and resistance. *Biochemistry*, 53(10):1565–1574, Mar 2014.

[102] H. Mosaei and J. Harbottle. Mechanisms of antibiotics inhibiting bacterial RNA polymerase. *Biochem. Soc. Trans.*, 47(1):339–350, Feb 2019.

[103] C. Ma, X. Yang, and P. J. Lewis. Bacterial Transcription as a Target for Antibacterial Drug Development. *Microbiol. Mol. Biol. Rev.*, 80(1):139–160, Mar 2016.

[104] R. E. Ashley, A. Dittmore, S. A. McPherson, C. L. Turnbough, K. C. Neuman, and N. Osheroff. Activities of gyrase and topoisomerase IV on positively supercoiled DNA. *Nucleic Acids Res.*, 45(16):9611–9624, Sep 2017.

[105] M. X. Ho, B. P. Hudson, K. Das, E. Arnold, and R. H. Ebright. Structures of RNA polymerase-antibiotic complexes. *Curr. Opin. Struct. Biol.*, 19(6):715–723, Dec 2009.

[106] S. Tuske, S. G. Sarafianos, X. Wang, B. Hudson, E. Sineva, J. Mukhopadhyay, J. J. Birktoft, O. Leroy, S. Ismail, A. D. Clark, C. Dharia, A. Napoli, O. Laptenko, J. Lee, S. Borukhov, R. H. Ebright, and E. Arnold. Inhibition of bacterial RNA polymerase by streptolydigin: stabilization of a straight-bridge-helix active-center conformation. *Cell*, 122(4):541–552, Aug 2005.

[107] G. A. Belogurov, M. N. Vassylyeva, A. Sevostyanova, J. R. Appleman, A. X. Xiang, R. Lira, S. E. Webber, S. Klyuyev, E. Nudler, I. Artsimovitch, and D. G. Vassylyev. Transcription inactivation through local refolding of the RNA polymerase structure. *Nature*, 457(7227):332–335, Jan 2009.

[108] C. J. Willmott, S. E. Critchlow, I. C. Eperon, and A. Maxwell. The complex of DNA gyrase and quinolone drugs with DNA forms a barrier to transcription by RNA polymerase. *J. Mol. Biol.*, 242(4):351–363, Sep 1994.

[109] B. S. Laursen, H. P. Sørensen, K. K. Mortensen, and H. U. Sperling-Petersen. Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.*, 69(1):101–123, Mar 2005.

[110] L. S. McCoy, Y. Xie, and Y. Tor. Antibiotics that target protein synthesis. *Wiley Interdiscip Rev RNA*, 2(2):209–232, 2011.

[111] A. Bermingham and J. P. Derrick. The folic acid biosynthesis pathway in bacteria: evaluation of potential for antibacterial drug discovery. *Bioessays*, 24(7):637–648, Jul 2002.

[112] A. Prasetyoputri, A. M. Jarrad, M. A. Cooper, and M. A. T. Blaskovich. The Eagle Effect and Antibiotic-Induced Persistence: Two Sides of the Same Coin? *Trends Microbiol.*, 27(4):339–354, Apr 2019.

[113] A. P. Magiorakos, A. Srinivasan, R. B. Carey, Y. Carmeli, M. E. Falagas, C. G. Giske, S. Harbarth, J. F. Hindler, G. Kahlmeter, B. Olsson-Liljequist, D. L. Paterson, L. B. Rice, J. Stelling, M. J. Struelens, A. Vatopoulos, J. T. Weber, and D. L. Monnet. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.*, 18(3):268–281, Mar 2012.

[114] J. M. Blair, M. A. Webber, A. J. Baylay, D. O. Ogbolu, and L. J. Piddock. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.*, 13(1):42–51, Jan 2015.

[115] J. M. Munita and C. A. Arias. Mechanisms of Antibiotic Resistance. *Microbiol Spectr*, 4(2), Apr 2016.

[116] W. Wang, Q. Guo, X. Xu, Z. K. Sheng, X. Ye, and M. Wang. High-level tetracycline resistance mediated by efflux pumps Tet(A) and Tet(A)-1 with two start codons. *J. Med. Microbiol.*, 63(Pt 11):1454–1459, Nov 2014.

[117] H. A. Terzi, C. Kulah, and I. H. Ciftci. The effects of active efflux pumps on antibiotic resistance in Pseudomonas aeruginosa. *World J. Microbiol. Biotechnol.*, 30(10):2681–2687, Oct 2014.

[118] K. Bush. Proliferation and significance of clinically relevant Î²-lactamases. *Ann. N. Y. Acad. Sci.*, 1277:84–90, Jan 2013.

[119] J. M. Rodriguez-Martinez, M. E. Cano, C. Velasco, L. Martinez-Martinez, and A. Pascual. Plasmid-mediated quinolone resistance: an update. *J. Infect. Chemother.*, 17(2):149–182, Apr 2011.

[120] J. Strahilevitz, G. A. Jacoby, D. C. Hooper, and A. Robicsek. Plasmid-mediated quinolone resistance: a multifaceted threat. *Clin. Microbiol. Rev.*, 22(4):664–689, Oct 2009.

[121] Idan Yelin and Roy Kishony. Antibiotic resistance. *Cell*, 172 (5):1136 – 1136.e1, 2018. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2018.02.018.

[122] M. Baym, T. D. Lieberman, E. D. Kelsic, R. Chait, R. Gross, I. Yelin, and R. Kishony. Spatiotemporal microbial evolution on antibiotic landscapes. *Science*, 353(6304):1147–1151, Sep 2016.

[123] N. A. Moran. Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, 108(5):583–586, Mar 2002.

[124] E. V. Koonin and Y. I. Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, 36(21):6688–6719, Dec 2008.

[125] Y. Tutar. Pseudogenes. *Comp. Funct. Genomics*, 2012:424526, 2012.

[126] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli. The microbial pan-genome. *Curr. Opin. Genet. Dev.*, 15(6):589–594, Dec 2005.

[127] L. Rouli, V. Merhej, P. E. Fournier, and D. Raoult. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect*, 7:72–85, Sep 2015.

[128] J. D. van Elsas, A. V. Semenov, R. Costa, and J. T. Trevors. Survival of Escherichia coli in the environment: fundamental and public health aspects. *ISME J*, 5(2):173–183, Feb 2011.

[129] P. Durao, R. Balbontin, and I. Gordo. Evolutionary Mechanisms Shaping the Maintenance of Antibiotic Resistance. *Trends Microbiol.*, 26(8):677–691, Aug 2018.

[130] E. Denamur and I. Matic. Evolution of mutation rates in bacteria. *Mol. Microbiol.*, 60(4):820–827, May 2006.

[131] H. Echols, C. Lu, and P. M. Burgers. Mutator strains of Escherichia coli, mutD and dnaQ, with defective exonucleolytic editing by DNA polymerase III holoenzyme. *Proc. Natl. Acad. Sci. U.S.A.*, 80(8):2189–2192, Apr 1983.

[132] H. Li, Y. F. Luo, B. J. Williams, T. S. Blackwell, and C. M. Xie. Structure and function of OprD protein in Pseudomonas aeruginosa: from antibiotic resistance to novel therapies. *Int. J. Med. Microbiol.*, 302(2):63–68, Mar 2012.

[133] A. Johnning, E. Kristiansson, J. Fick, B. Weijdegard, and D. G. Larsson. Resistance Mutations in gyrA and parC are Common in Escherichia Communities of both Fluoroquinolone-Polluted and Uncontaminated Aquatic Environments. *Front Microbiol*, 6: 1355, 2015.

[134] E. Darmon and D. R. Leach. Bacterial genome instability. *Microbiol. Mol. Biol. Rev.*, 78(1):1–39, Mar 2014.

[135] V. Periwal and V. Scaria. Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics*, 31(1):1–9, Jan 2015.

[136] M. G. Kidwell. *Encyclopedia of Genetics*, chapter Horizontal Transfer, pages 973–975. Academic Press, New York, 2001. ISBN 978-0-12-227080-2. doi: https://doi.org/10.1006/rwgn.2001.0632.

[137] O. Popa, E. Hazkani-Covo, G. Landan, W. Martin, and T. Dagan. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.*, 21(4):599–609, Apr 2011.

[138] C. M. Thomas and K. M. Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.*, 3(9):711–721, Sep 2005.

[139] L. S. Frost, R. Leplae, A. O. Summers, and A. Toussaint. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, 3(9):722–732, Sep 2005.

[140] T. Vigil-Stenman, K. Ininbergs, B. Bergman, and M. Ekman. High abundance and expression of transposases in bacteria from the Baltic Sea. *ISME J*, 11(11):2611–2623, Nov 2017.

[141] A. Orlek, H. Phan, A. E. Sheppard, M. Doumith, M. Ellington, T. Peto, D. Crook, A. S. Walker, N. Woodford, M. F. Anjum, and N. Stoesser. Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid*, 91:42–52, May 2017.

[142] F. Dionisio, R. Zilhao, and J. A. Gama. Interactions between plasmids and other mobile genetic elements affect their transmission and persistence. *Plasmid*, 102:29–36, Mar 2019.

[143] C. Smillie, M. P. Garcillan-Barcia, M. V. Francia, E. P. Rocha, and F. de la Cruz. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, 74(3):434–452, Sep 2010.

[144] Y. Y. Liu, Y. Wang, T. R. Walsh, L. X. Yi, R. Zhang, J. Spencer, Y. Doi, G. Tian, B. Dong, X. Huang, L. F. Yu, D. Gu, H. Ren, X. Chen, L. Lv, D. He, H. Zhou, Z. Liang, J. H. Liu, and J. Shen. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect Dis*, 16(2): 161–168, Feb 2016.

[145] R. Wang, L. van Dorp, L. P. Shaw, P. Bradley, Q. Wang, X. Wang, L. Jin, Q. Zhang, Y. Liu, A. Rieux, T. Dorai-Schneiders, L. A. Weinert, Z. Iqbal, X. Didelot, H. Wang, and F. Balloux. The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat Commun*, 9(1):1179, Mar 2018.

[146] B. B. Xavier, C. Lammens, R. Ruhal, S. Kumar-Singh, P. Butaye, H. Goossens, and S. Malhotra-Kumar. Identification of a novel plasmid-mediated colistin-resistance gene, mcr-2, in Escherichia coli, Belgium, June 2016. *Euro Surveill.*, 21(27), Jul 2016.

[147] W. Yin, H. Li, Y. Shen, Z. Liu, S. Wang, Z. Shen, R. Zhang, T. R. Walsh, J. Shen, and Y. Wang. Novel Plasmid-Mediated Colistin Resistance Gene mcr-3 in Escherichia coli. *MBio*, 8(3), Jun 2017.

[148] A. Carattoli, L. Villa, C. Feudi, L. Curcio, S. Orsini, A. Luppi, G. Pezzotti, and C. F. Magistrali. Novel plasmid-mediated colistin resistance mcr-4 gene in Salmonella and Escherichia coli, Italy 2013, Spain and Belgium, 2015 to 2016. *Euro Surveill.*, 22(31), Aug 2017.

[149] M. Borowiak, J. Fischer, J. A. Hammerl, R. S. Hendriksen, I. Szabo, and B. Malorny. Identification of a novel transposon-associated phosphoethanolamine transferase gene, mcr-5, conferring colistin resistance in d-tartrate fermenting Salmonella enterica subsp. enterica serovar Paratyphi B. *J. Antimicrob. Chemother.*, 72(12):3317–3324, Dec 2017.

[150] M. AbuOun, E. J. Stubberfield, N. A. Duggett, M. Kirchner, L. Dormer, J. Nunez-Garcia, L. P. Randall, F. Lemma, D. W. Crook, C. Teale, R. P. Smith, and M. F. Anjum. mcr-1 and mcr-2 variant genes identified in Moraxella species isolated from pigs in Great Britain from 2014 to 2015. *J. Antimicrob. Chemother.*, 72 (10):2745–2749, Oct 2017.

[151] Y. Q. Yang, Y. X. Li, C. W. Lei, A. Y. Zhang, and H. N. Wang. Novel plasmid-mediated colistin resistance gene mcr-7.1 in Klebsiella pneumoniae. *J. Antimicrob. Chemother.*, 73(7):1791–1795, Jul 2018.

[152] X. Wang, Y. Wang, Y. Zhou, J. Li, W. Yin, S. Wang, S. Zhang, J. Shen, Z. Shen, and Y. Wang. Emergence of a novel mobile colistin resistance gene, mcr-8, in NDM-producing Klebsiella pneumoniae. *Emerg Microbes Infect*, 7(1):122, Jul 2018.

[153] L. M. Carroll, A. Gaballa, C. Guldimann, G. Sullivan, L. O. Henderson, and M. Wiedmann. Identification of Novel Mobilized Colistin Resistance Gene mcr-9 in a Multidrug-Resistant, Colistin-Susceptible Salmonella enterica Serotype Typhimurium Isolate. *MBio*, 10(3), May 2019.

[154] N. Waglechner and G. D. Wright. Antibiotic resistance: it's bad, but why isn't it worse? *BMC Biol.*, 15(1):84, Sep 2017.

[155] R. Laxminarayan, C. F. Amabile-Cuevas, O. Cars, T. Evans, D. L. Heymann, S. Hoffman, A. Holmes, M. Mendelson, D. Sridhar, M. Woolhouse, and J. A. Røttingen. UN High-Level Meeting on antimicrobials – what do we need? *Lancet*, 388(10041):218–220, Jul 2016.

[156] L. B. Rice. Progress and challenges in implementing the research on ESKAPE pathogens. *Infect Control Hosp Epidemiol*, 31 Suppl 1:7–10, Nov 2010.

[157] Centers for Disease Control and Prevention (CDC). Antibiotic Resistance Threats in the United States, 2013. URL `https://www.cdc.gov/drugresistance/threat-report-2013/pdf/ar-threats-2013-508.pdf`. Accessed: 14.06.2019.

[158] World Health Organization (WHO). Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics, Feb 2017. URL https://www.who.int/medicines/publications/WHO-PPL-Short_Summary_25Feb-ET_NM_WHO.pdf. Accessed: 14.06.2019.

[159] J. P. Burnham, M. A. Olsen, and M. H. Kollef. Re-estimating annual deaths due to multidrug-resistant organism infections. *Infect Control Hosp Epidemiol*, 40(1):112–113, Jan 2019.

[160] J. O'Neill. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations, Dec 2014. URL https://amr-review.org/sites/default/files/AMR%20Review%20Paper%20-%20Tackling%20a%20crisis%20for%20the%20health%20and%20wealth%20of%20nations_1.pdf. Accessed: 14.06.2019.

[161] T. Jirka, M. Hafner, E. Yerushalmi, R. Smith, J. Bellasio, R. Vardavas, T. Bienkowska-Gibbs, and J. Rubin. Estimating the economic costs of antimicrobial resistance: Model and Results, 2014. URL https://www.rand.org/pubs/research_reports/RR911.html. Accessed: 14.06.2019.

[162] J. A. Ayukekbong, M. Ntemgwa, and A. N. Atabe. The threat of antimicrobial resistance in developing countries: causes and control strategies. *Antimicrob Resist Infect Control*, 6:47, 2017.

[163] A. H. Holmes, L. S. Moore, A. Sundsfjord, M. Steinbakk, S. Regmi, A. Karkey, P. J. Guerin, and L. J. Piddock. Understanding the mechanisms and drivers of antimicrobial resistance. *Lancet*, 387(10014):176–187, Jan 2016.

[164] U. Theuretzbacher, S. Gottwalt, P. Beyer, M. Butler, L. Czaplewski, C. Lienhardt, L. Moja, M. Paul, S. Paulin, J. H. Rex, L. L. Silver, M. Spigelman, G. E. Thwaites, J. P. Paccaud, and S. Harbarth. Analysis of the clinical antibacterial and antituberculosis pipeline. *Lancet Infect Dis*, 19(2):e40–e50, Feb 2019.

[165] Boston Consulting Group (BCG). Vaccines to tackle drug resistant infections: An evaluation of R&D opportunities, 2018. URL https://vaccinesforamr.org/wp-content/uploads/2018/09/Vaccines_for_AMR.pdf. Accessed: 14.06.2019.

[166] M. C. Danovaro-Holliday, S. Garcia, C. de Quadros, G. Tambini, and J. K. Andrus. Progress in vaccination against Haemophilus influenzae type b in the Americas. *PLoS Med.*, 5(4):e87, Apr 2008.

[167] C. A. MacLennan, L. B. Martin, and F. Micoli. Vaccines against invasive Salmonella disease: current status and future directions. *Hum Vaccin Immunother*, 10(6):1478–1493, 2014.

[168] G. L. Kim, S. H. Seon, and D. K. Rhee. Pneumonia and Streptococcus pneumoniae vaccine. *Arch. Pharm. Res.*, 40(8):885–893, Aug 2017.

[169] J. C. Kwong, S. Maaten, R. E. Upshur, D. M. Patrick, and F. Marra. The effect of universal influenza immunization on antibiotic prescriptions: an ecological study. *Clin. Infect. Dis.*, 49(5):750–756, Sep 2009.

[170] M. S. Butler, M. A. Blaskovich, and M. A. Cooper. Antibiotics in the clinical pipeline in 2013. *J. Antibiot.*, 66(10):571–591, Oct 2013.

[171] M. Tyers and G. D. Wright. Drug combinations: a strategy to extend the life of antibiotics in the 21st century. *Nat. Rev. Microbiol.*, 17(3):141–155, Mar 2019.

[172] D. R. Roach, C. Y. Leung, M. Henry, E. Morello, D. Singh, J. P. Di Santo, J. S. Weitz, and L. Debarbieux. Synergy between the Host Immune System and Bacteriophage Is Essential for Successful Phage Therapy against an Acute Respiratory Pathogen. *Cell Host Microbe*, 22(1):38–47, Jul 2017.

[173] J. M. Sweere, J. D. Van Belleghem, H. Ishak, M. S. Bach, M. Popescu, V. Sunkari, G. Kaber, R. Manasherob, G. A. Suh, X. Cao, C. R. de Vries, D. N. Lam, P. L. Marshall, M. Birukova, E. Katznelson, D. V. Lazzareschi, S. Balaji, S. G. Keswani, T. R. Hawn, P. R. Secor, and P. L. Bollyky. Bacteriophage trigger antiviral immunity and prevent clearance of bacterial infection. *Science*, 363(6434), Mar 2019.

[174] K. E. Kortright, B. K. Chan, J. L. Koff, and P. E. Turner. Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host Microbe*, 25(2):219–232, Feb 2019.

[175] J. L. Fox. Antimicrobial peptides stage a comeback. *Nat. Biotechnol.*, 31(5):379–382, May 2013.

[176] Z. Wang and G. Wang. APD: the Antimicrobial Peptide Database. *Nucleic Acids Res.*, 32(Database issue):D590–592, Jan 2004.

[177] C. D. Fjell, J. A. Hiss, R. E. Hancock, and G. Schneider. Designing antimicrobial peptides: form follows function. *Nat Rev Drug Discov*, 11(1):37–51, Dec 2011.

[178] P. D. Cotter, R. P. Ross, and C. Hill. Bacteriocins – a viable alternative to antibiotics? *Nat. Rev. Microbiol.*, 11(2):95–105, Feb 2013.

[179] C. Saylor, E. Dadachova, and A. Casadevall. Monoclonal antibody-based therapies for microbial diseases. *Vaccine*, 27 Suppl 6:38–46, Dec 2009.

[180] E. Nagy, G. Nagy, C. A. Power, A. Badarau, and V. Szijártó. *Recombinant Antibodies for Infectious Diseases*, chapter "Antibacterial Monoclonal Antibodies", pages 119–153. Springer International Publishing, Cham, 2017. ISBN 978-3-319-72077-7. doi: 10.1007/978-3-319-72077-7_7.

[181] M. N. Ragheb, M. K. Thomason, C. Hsu, P. Nugent, J. Gage, A. N. Samadpour, A. Kariisa, C. N. Merrikh, S. I. Miller, D. R. Sherman, and H. Merrikh. Inhibiting the Evolution of Antibiotic Resistance. *Mol. Cell*, 73(1):157–165, Jan 2019.

[182] K. French, J. Evans, H. Tanner, S. Gossain, and A. Hussain. The Clinical Impact of Rapid, Direct MALDI-ToF Identification of Bacteria from Positive Blood Cultures. *PLoS ONE*, 11(12): e0169332, 2016.

[183] A. Akram, M. Maley, I. Gosbell, T. Nguyen, and R. Chavada. Utility of 16S rRNA PCR performed on clinical specimens in patient management. *Int. J. Infect. Dis.*, 57:144–149, Apr 2017.

[184] C. R. Cox, K. R. Jensen, N. R. Saichek, and K. J. Voorhees. Strain-level bacterial identification by CeO2-catalyzed MALDI-TOF MS fatty acid analysis and comparison to commercial protein-based methods. *Sci Rep*, 5:10470, Jul 2015.

[185] F. J. May, R. J. Stafford, H. Carroll, J. M. Robson, R. Vohra, G. R. Nimmo, J. Bates, M. D. Kirk, E. J. Fearnley, and B. G. Polkinghorne. The effects of culture independent diagnostic testing on the diagnosis and reporting of enteric bacterial pathogens in Queensland, 2010 to 2014. *Commun Dis Intell Q Rep*, 41(3): E223–E230, Sep 2017.

[186] N. L. Bachmann, R. J. Rockett, V. J. Timms, and V. Sintchenko. Advances in Clinical Sample Preparation for Identification and Characterization of Bacterial Pathogens Using Metagenomics. *Front Public Health*, 6:363, 2018.

[187] G. Langley, J. Besser, M. Iwamoto, F. C. Lessa, A. Cronquist, T. H. Skoff, S. Chaves, D. Boxrud, R. W. Pinner, and L. H. Harrison. Effect of Culture-Independent Diagnostic Tests on Future Emerging Infections Program Surveillance. *Emerging Infect. Dis.*, 21(9):1582–1588, Sep 2015.

[188] P. E. Fournier, G. Dubourg, and D. Raoult. Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Med*, 6(11):114, 2014.

[189] T. A. Blauwkamp, S. Thair, M. J. Rosen, L. Blair, M. S. Lindner, I. D. Vilfan, T. Kawli, F. C. Christians, S. Venkatasubrahmanyam, G. D. Wall, A. Cheung, Z. N. Rogers, G. Meshulam-Simon, L. Huijse, S. Balakrishnan, J. V. Quinn, D. Hollemon, D. K. Hong, M. L. Vaughn, M. Kertesz, S. Bercovici, J. C. Wilber, and

S. Yang. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat Microbiol*, 4(4):663–674, Apr 2019.

[190] C. Langelier, K. L. Kalantar, F. Moazed, M. R. Wilson, E. D. Crawford, T. Deiss, A. Belzer, S. Bolourchi, S. Caldera, M. Fung, A. Jauregui, K. Malcolm, A. Lyden, L. Khan, K. Vessel, J. Quan, M. Zinter, C. Y. Chiu, E. D. Chow, J. Wilson, S. Miller, M. A. Matthay, K. S. Pollard, S. Christenson, C. S. Calfee, and J. L. DeRisi. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. *Proc. Natl. Acad. Sci. U.S.A.*, 115(52):E12353–E12362, Dec 2018.

[191] R. Schlaberg, C. Y. Chiu, S. Miller, G. W. Procop, and G. Weinstock. Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. *Arch. Pathol. Lab. Med.*, 141(6):776–786, Jun 2017.

[192] J. Lloyd-Price, G. Abu-Ali, and C. Huttenhower. The healthy human microbiome. *Genome Med*, 8(1):51, Apr 2016.

[193] European Centre for Disease Prevention and Control (ECDC). ECDC calls for continued action to address antimicrobial resistance in healthcare settings, Nov 2018. URL `https://ecdc.europa.eu/en/news-events/ecdc-calls-continued-action-address-antimicrobial-resistance-healthcare-settings`. Accessed: 17.06.2019.

[194] World Health Organization (WHO). 2015, 2015. URL `https://www.who.int/glass/en/`. Accessed: 17.06.2019.

[195] European Commission (EC). A European One Health Action Plan against Antimicrobial Resistance (AMR), Jun 2017. URL `https://ec.europa.eu/health/amr/sites/amr/files/amr_action_plan_2017_en.pdf`. Accessed: 17.06.2019.

[196] Centers for Disease Control and Prevention (CDC). National action plan for combating antibiotic-resistant bacteria, Mar 2015. URL `https://www.cdc.gov/drugresistance/pdf/national_action_plan_for_combating_antibotic-resistant_bacteria.pdf`. Accessed: 17.06.2019.

[197] M. R. Farhat, L. Freschi, R. Calderon, T. Ioerger, M. Snyder, C. J. Meehan, B. de Jong, L. Rigouts, A. Sloutsky, D. Kaur, S. Sunyaev, D. van Soolingen, J. Shendure, J. Sacchettini, and M. Murray. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat Commun*, 10(1):2128, May 2019.

[198] R. M. Martin, J. Cao, W. Wu, L. Zhao, D. M. Manthei, A. Pirani, E. Snitkin, P. N. Malani, K. Rao, and M. A. Bachman. Identification of Pathogenicity-Associated Loci in Klebsiella pneumoniae from Hospitalized Patients. *mSystems*, 3(3), 2018.

[199] H. Rohde, J. Qin, Y. Cui, D. Li, N. J. Loman, M. Hentschke, W. Chen, F. Pu, Y. Peng, J. Li, F. Xi, S. Li, Y. Li, Z. Zhang, X. Yang, M. Zhao, P. Wang, Y. Guan, Z. Cen, X. Zhao, M. Christner, R. Kobbe, S. Loos, J. Oh, L. Yang, A. Danchin, G. F. Gao, Y. Song, Y. Li, H. Yang, J. Wang, J. Xu, M. J. Pallen, J. Wang, M. Aepfelbacher, R. Yang, K. E. Holt, D. J. Studholme, M. Feldgarden, and M. Manrique. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N. Engl. J. Med.*, 365(8):718–724, Aug 2011.

[200] S. A. Ahmed, J. Awosika, C. Baldwin, K. A. Bishop-Lilly, B. Biswas, S. Broomall, P. S. Chain, O. Chertkov, O. Chokoshvili, S. Coyne, K. Davenport, J. C. Detter, W. Dorman, T. H. Erkkila, J. P. Folster, K. G. Frey, M. George, C. Gleasner, M. Henry, K. K. Hill, K. Hubbard, J. Insalaco, S. Johnson, A. Kitzmiller, M. Krepps, C. C. Lo, T. Luu, L. A. McNew, T. Minogue, C. A. Munk, B. Osborne, M. Patel, K. G. Reitenga, C. N. Rosenzweig, A. Shea, X. Shen, N. Strockbine, C. Tarr, H. Teshima, E. van Gieson, K. Verratti, M. Wolcott, G. Xie, S. Sozhamannan, and H. S. Gibbons. Genomic comparison of Escherichia coli O104:H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including shiga toxin encoding phage stx2. *PLoS ONE*, 7(11):e48228, 2012.

[201] A. C. Schurch and W. van Schaik. Challenges and opportunities for whole-genome sequencing-based surveillance of antibiotic resistance. *Ann. N. Y. Acad. Sci.*, 1388(1):108–120, Jan 2017.

[202] F. P. Maurer, M. Christner, M. Hentschke, and H. Rohde. Advances in Rapid Identification and Susceptibility Testing of Bacteria in the Clinical Microbiology Laboratory: Implications for Patient Care and Antimicrobial Stewardship Programs. *Infect Dis Rep*, 9(1):6839, Mar 2017.

[203] A. R. Wattam, D. Abraham, O. Dalay, T. L. Disz, T. Driscoll, J. L. Gabbard, J. J. Gillespie, R. Gough, D. Hix, R. Kenyon, D. Machi, C. Mao, E. K. Nordberg, R. Olson, R. Overbeek, G. D. Pusch, M. Shukla, J. Schulman, R. L. Stevens, D. E. Sullivan, V. Vonstein, A. Warren, R. Will, M. J. Wilson, H. S. Yoo, C. Zhang, Y. Zhang, and B. W. Sobral. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, 42(Database issue): D581–591, Jan 2014.

[204] J. Wang and H. Jia. Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.*, 14(8):508–522, Aug 2016.

[205] G. M. Weinstock. Genomic approaches to studying the human microbiota. *Nature*, 489(7415):250–256, Sep 2012.

[206] C. S. Riesenfeld, P. D. Schloss, and J. Handelsman. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, 38:525–552, 2004.

[207] P. Yarza, P. Yilmaz, E. Pruesse, F. O. Glockner, W. Ludwig, K. H. Schleifer, W. B. Whitman, J. Euzeby, R. Amann, and R. Rossello-Mora. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.*, 12(9):635–645, Sep 2014.

[208] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neubock, and I. L. Hofacker. The Vienna RNA websuite. *Nucleic Acids Res.*, 36 (Web Server issue):W70–74, Jul 2008.

[209] P. Kerpedjiev, S. Hammer, and I. L. Hofacker. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, 31(20):3377–3379, Oct 2015.

[210] F. Teng, S. S. Darveekaran Nair, P. Zhu, S. Li, S. Huang, X. Li, J. Xu, and F. Yang. Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. *Sci Rep*, 8(1):16321, Nov 2018.

[211] J. M. Janda and S. L. Abbott. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.*, 45(9):2761–2764, Sep 2007.

[212] T. Coenye and P. Vandamme. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol. Lett.*, 228(1):45–49, Nov 2003.

[213] W. Pootakham, W. Mhuantong, T. Yoocha, L. Putchim, C. Sonthirod, C. Naktang, N. Thongtham, and S. Tangphatsornruang. High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Sci Rep*, 7(1):2774, Jun 2017.

[214] N. P. Nguyen, T. Warnow, M. Pop, and B. White. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microbiomes*, 2:16004, 2016.

[215] S. Sathiananthamoorthy, J. Malone-Lee, K. Gill, A. Tymon, T. K. Nguyen, S. Gurung, L. Collins, A. S. Kupelian, S. Swamy, R. Khasriya, D. A. Spratt, and J. L. Rohn. Reassessment of Routine Midstream Culture in Diagnosis of Urinary Tract Infection. *J. Clin. Microbiol.*, 57(3), Mar 2019.

[216] P. D. Schloss, M. L. Jenior, C. C. Koumpouras, S. L. Westcott, and S. K. Highlander. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ*, 4:e1869, 2016.

[217] J. Wagner, P. Coupland, H. P. Browne, T. D. Lawley, S. C. Francis, and J. Parkhill. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol.*, 16(1): 274, Nov 2016.

[218] A. Benitez-Paez, K. J. Portune, and Y. Sanz. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinIONâ„¢ portable nanopore sequencer. *Gigascience*, 5:4, 2016.

[219] J. A. Klappenbach, J. M. Dunbar, and T. M. Schmidt. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.*, 66(4):1328–1333, Apr 2000.

[220] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, 35(9):833–844, Sep 2017.

[221] D. D. Kang, J. Froula, R. Egan, and Z. Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 2015.

[222] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nat. Methods*, 11(11):1144–1146, Nov 2014.

[223] M. Strous, B. Kraft, R. Bisdorf, and H. E. Tegetmeyer. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol*, 3:410, 2012.

[224] F. Meyer, P. Hofmann, P. Belmann, R. Garrido-Oter, A. Fritz, A. Sczyrba, and A. C. McHardy. AMBER: Assessment of Metagenome BinnERs. *Gigascience*, 7(6), Jun 2018.

[225] R. R. Miller, V. Montoya, J. L. Gardy, D. M. Patrick, and P. Tang. Metagenomics for pathogen detection in public health. *Genome Med*, 5(9):81, 2013.

[226] S. Amrane and J.-C. Lagier. Metagenomic and clinical microbiology. *Human Microbiome Journal*, 9:1–6, 2018.

[227] N. J. Loman, C. Constantinidou, M. Christner, H. Rohde, J. Z. Chan, J. Quick, J. C. Weir, C. Quince, G. P. Smith, J. R. Betley, M. Aepfelbacher, and M. J. Pallen. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *JAMA*, 309(14):1502–1510, Apr 2013.

[228] R. S. Hendriksen, P. Munk, P. Njage, B. van Bunnik, L. McNally, O. Lukjancenko, T. Roder, D. Nieuwenhuijse, S. K. Pedersen, J. Kjeldgaard, R. S. Kaas, P. T. L. C. Clausen, J. K. Vogt, P. Leekitcharoenphon, M. G. M. van de Schans, T. Zuidema, A. M. de Roda Husman, S. Rasmussen, B. Petersen, C. Amid, G. Cochrane, T. Sicheritz-Ponten, H. Schmitt, J. R. M. Alvarez, A. Aidara-Kane, S. J. Pamp, O. Lund, T. Hald, M. Woolhouse, M. P. Koopmans, H. Vigre, T. N. Petersen, F. M. Aarestrup, A. Bego, C. Rees, S. Cassar, K. Coventry, P. Collignon, F. Allerberger, T. O. Rahube, G. Oliveira, I. Ivanov, Y. Vuthy, T. Sopheak,

C. K. Yost, C. Ke, H. Zheng, L. Baisheng, X. Jiao, P. Donado-Godoy, K. J. Coulibaly, M. Jergović, J. Hrenovic, R. Karpíšková, J. E. Villacis, M. Legesse, T. Eguale, A. Heikinheimo, L. Malania, A. Nitsche, A. Brinkmann, C. K. S. Saba, B. Kocsis, N. Solymosi, T. R. Thorsteinsdottir, A. M. Hatha, M. Alebouyeh, D. Morris, M. Cormican, L. O'Connor, J. Moran-Gilad, P. Alba, A. Battisti, Z. Shakenova, C. Kiiyukia, E. Ng'eno, L. Raka, J. Avsejenko, A. Bērziņš, V. Bartkevics, C. Penny, H. Rajandas, S. Parimannan, M. V. Haber, P. Pal, G. J. Jeunen, N. Gemmell, K. Fashae, R. Holmstad, R. Hasan, S. Shakoor, M. L. Z. Rojas, D. Wasyl, G. Bosevska, M. Kochubovski, C. Radu, A. Gassama, V. Radosavljevic, S. Wuertz, R. Zuniga-Montanez, M. Y. F. Tay, D. Gavačová, K. Pastuchova, P. Truska, M. Trkov, K. Esterhuyse, K. Keddy, M. Cerdà-Cuéllar, S. Pathirage, L. Norrgren, S. Örn, D. G. J. Larsson, T. V. Heijden, H. H. Kumburu, B. Sanneh, P. Bidjada, B. M. Njanpop-Lafourcade, S. C. Nikiema-Pessinaba, B. Levent, J. S. Meschke, N. K. Beck, C. D. Van, N. D. Phuc, D. M. N. Tran, G. Kwenda, D. A. Tabo, A. L. Wester, and S. Cuadros-Orellana. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat Commun*, 10(1):1124, Mar 2019.

[229] M. Roosaare, M. Puustusmaa, M. Mols, M. Vaher, and M. Remm. PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. *PeerJ*, 6:e4588, 2018.

[230] R. Rozov, A. Brown Kav, D. Bogumil, N. Shterzer, E. Halperin, I. Mizrahi, and R. Shamir. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics*, 33(4): 475–482, Feb 2017.

[231] D. Antipov, N. Hartwick, M. Shen, M. Raiko, A. Lapidus, and P. A. Pevzner. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*, 32(22):3380–3387, Nov 2016.

[232] F. Zhou and Y. Xu. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, 26(16):2051–2052, Aug 2010.

[233] E. Carloni, F. Andreoni, E. Omiccioli, L. Villa, M. Magnani, and A. Carattoli. Comparative analysis of the standard PCR-Based Replicon Typing (PBRT) with the commercial PBRT-KIT. *Plasmid*, 90:10–14, Mar 2017.

[234] A. Carattoli, E. Zankari, A. Garcia-Fernandez, M. Voldby Larsen, O. Lund, L. Villa, F. Møller Aarestrup, and H. Hasman. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.*, 58(7):3895–3903, Jul 2014.

[235] J. Robertson and J. H. E. Nash. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom*, 4(8), Aug 2018.

[236] M. Jaillard, L. Lima, M. Tournoud, P. Mahe, A. van Belkum, V. Lacroix, and L. Jacob. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet.*, 14(11):e1007758, Nov 2018.

[237] E. Aun, A. Brauer, V. Kisand, T. Tenson, and M. Remm. A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput. Biol.*, 14(10):e1006434, Oct 2018.

[238] J. A. Lees, M. Vehkala, N. Valimaki, S. R. Harris, C. Chewapreecha, N. J. Croucher, P. Marttinen, M. R. Davies, A. C. Steer, S. Y. Tong, A. Honkela, J. Parkhill, S. D. Bentley, and J. Corander. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun*, 7:12797, Sep 2016.

[239] J. J. Davis, S. Boisvert, T. Brettin, R. W. Kenyon, C. Mao, R. Olson, R. Overbeek, J. Santerre, M. Shukla, A. R. Wattam, R. Will, F. Xia, and R. Stevens. Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci Rep*, 6:27930, Jun 2016.

[240] L. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014.

[241] L. McInnes, J. Healy, N. Saul, and L. Großberger. UMAP: uniform manifold approximation and projection. *J. Open Source Software*, 3(29):861, 2018.

[242] B. Jia, A. R. Raphenya, B. Alcock, N. Waglechner, P. Guo, K. K. Tsang, B. A. Lago, B. M. Dave, S. Pereira, A. N. Sharma, S. Doshi, M. Courtot, R. Lo, L. E. Williams, J. G. Frye, T. Elsayegh, D. Sardar, E. L. Westman, A. C. Pawlowski, T. A. Johnson, F. S. Brinkman, G. D. Wright, and A. G. McArthur. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, 45(D1): D566–D573, Jan 2017.

[243] B. Liu, D. Zheng, Q. Jin, L. Chen, and J. Yang. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, 47(D1):D687–D692, Jan 2019.

[244] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 17(1):132, Jun 2016.

[245] A. Orlek, N. Stoesser, M. F. Anjum, M. Doumith, M. J. Ellington, T. Peto, D. Crook, N. Woodford, A. S. Walker, H. Phan, and A. E.

Sheppard. Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. *Front Microbiol*, 8:182, 2017.

[246] E. Garrison, J. Siren, A. M. Novak, G. Hickey, J. M. Eizenga, E. T. Dawson, W. Jones, S. Garg, C. Markello, M. F. Lin, B. Paten, and R. Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.*, 36(9):875–879, Oct 2018.

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Dr. Andreas Keller for giving me the opportunity to join his group, Chair for Clinical Bioinformatics (CCB) at the Saarland University, and for his continued support of my Ph.D. studies and the related research.

Furthermore, I would like to thank the rest of my thesis committee for their insightful comments and evaluation of my thesis.

I would also like to show my gratitude to the employees of Siemens AG, Curest GmbH, and Ares Genetics GmbH for working together with me on the GEAR project which became the main part of my study. Special thanks go to our industrial collaborators Dr. Achm Plum, and Dr. Andreas E. Posch, Dr. Susanne Schmolke, Dr. Cord Stähler, Dr. Howard Li, Dr. Laura Smoot, our associates at the University Hospital Homburg, Prof. Dr. Markus Bischoff, Prof. Dr. Lutz von Müller and Prof. Dr. Mathias Herrmann, and our associates at the Institute of Clinical Molecular Biology in Kiel Prof. Dr. Andre Franke and Dr. Georg-Hemmrich Stanisak.

Thanks are also due to Prof. Dr. Andreas Keller, Prof. Dr. Eckart Meese, and Prof. Dr. Dr. Thomas Lengauer for the funding of my Ph.D. work.

I would like to thank Prof. Dr. Nico Pfeifer, Prof. Dr. Tobias Marschall, Prof. Dr. Eckart Meese, Prof. Dr. Rolf Müller, Prof. Dr. Olga Kalinina and Prof. Dr. Mathias Hein for their helpful advice regarding different parts of my research.

My sincere thanks go to Dr. Christina Backes and Dr. Cédric C. Lazny for their tremendous support during my Ph.D. Their help and guidance has had a significant impact on my research. I am extremely grateful for their encouragement and moral support during tough times. I will also fondly remember my colleagues Tobias Fehlmann, Mustafa Kahraman, and Fabian Kern, our insightful discussions and the great time we had both during and outside of work. In addition, I would like to thank our secretary Sabine Lessel, along with past and present members of the CCB group.

I am grateful to my fellow Ph.D. students from the MPI, MMCI and CBI, and also to other friends from the Saarland University for all the fun we have had over the last few years.

Last but not least, I would like to thank my family for supporting me throughout my studies and life in general.