



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>

1672-0229 © 2019 The Authors. Production and hosting by Elsevier B.V. on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Please cite this article as: V. Galata, C. C. Laczny, C. Backes et al., Integrating Culture-based Antibiotic Resistance Profiles with Whole-genome Sequencing Data for 11,087 Clinical Isolates, Genomics Proteomics Bioinformatics, <https://doi.org/10.1016/j.gpb.2018.11.002>

Received 4 June 2018; revised 28 September 2018; accepted 7 November 2018
Available online xxx

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 2A). 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* ($\geq 29\%$ 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,

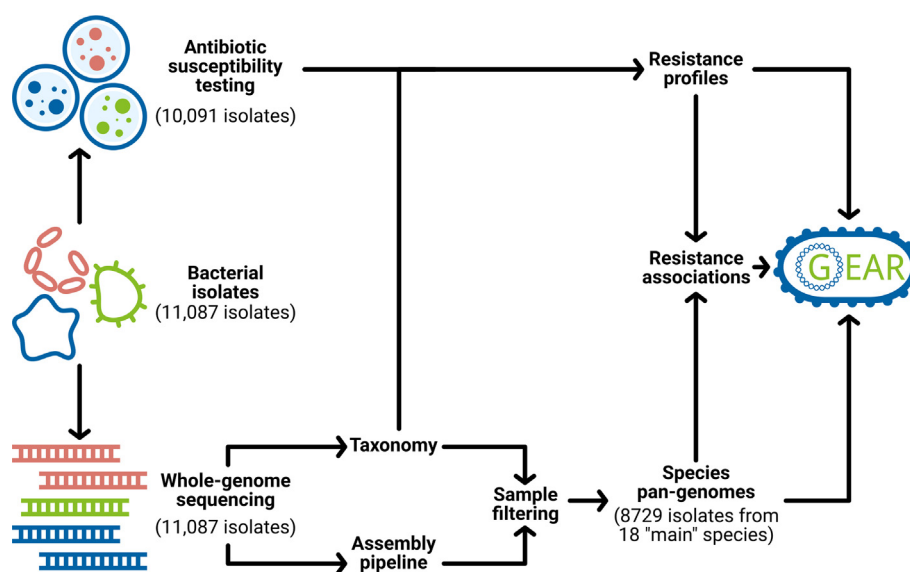


Figure 1 GEAR-base workflow and structure

Schematic overview of data collection, processing and integration into GEAR-base.

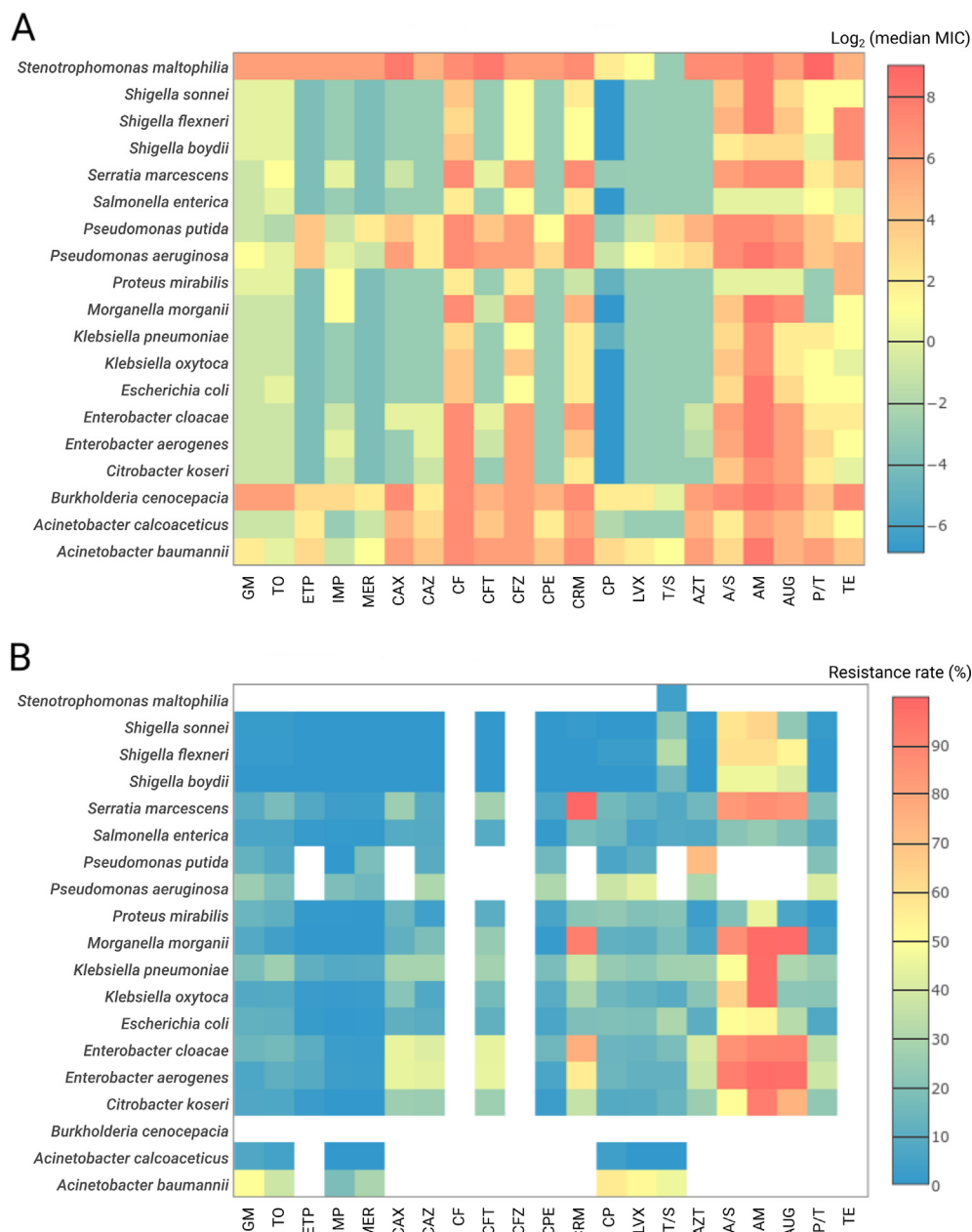


Figure 2 Overview of resistance profiles

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, result-

ing 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 marker 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

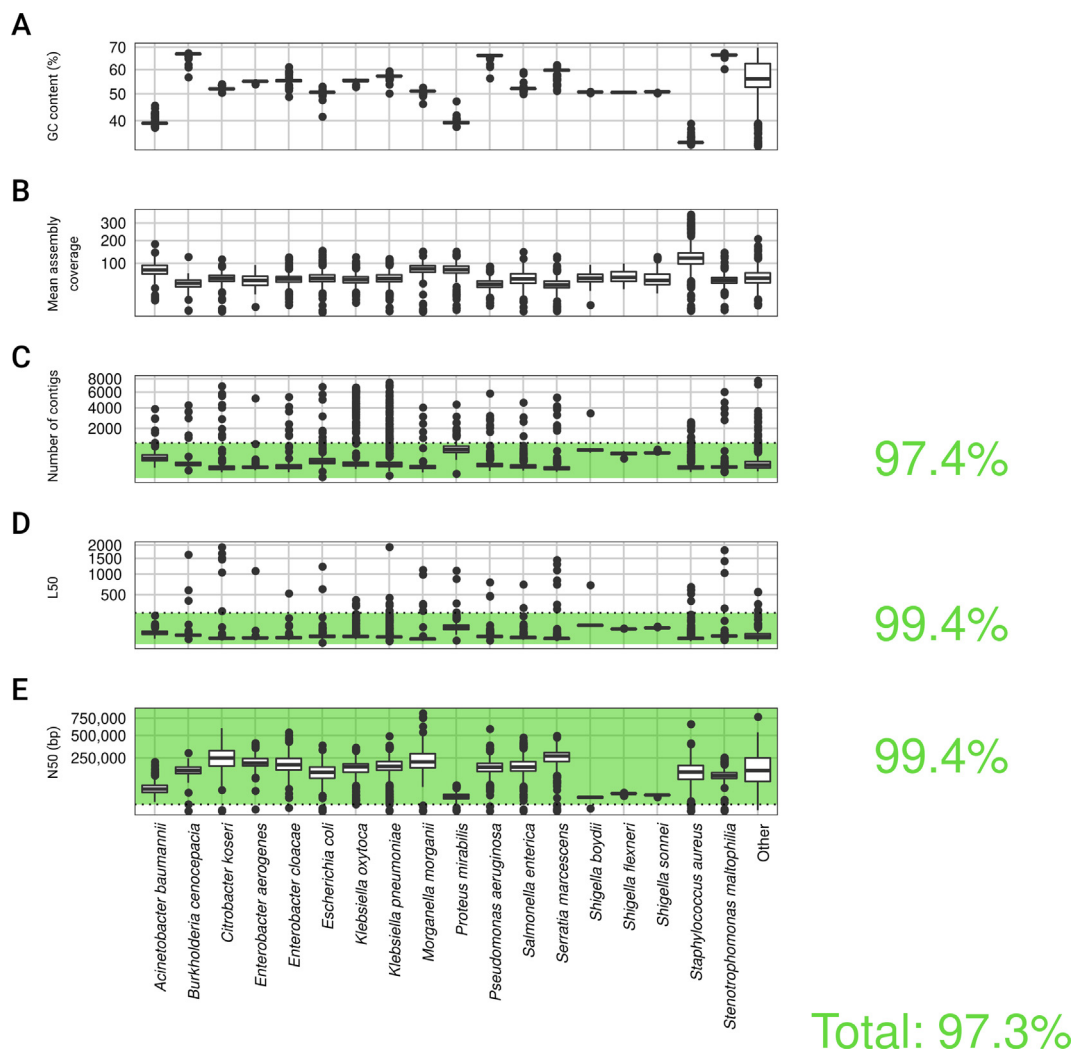


Figure 3 Assembly quality overview

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 ($\geq 10\%$) for *B. ceno-*

cepacia, *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 $\geq 90\%$ pair-wise amino acid sequence identity (Methods). Most centroids were found in $< 10\%$ or in $\geq 90\%$ 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.

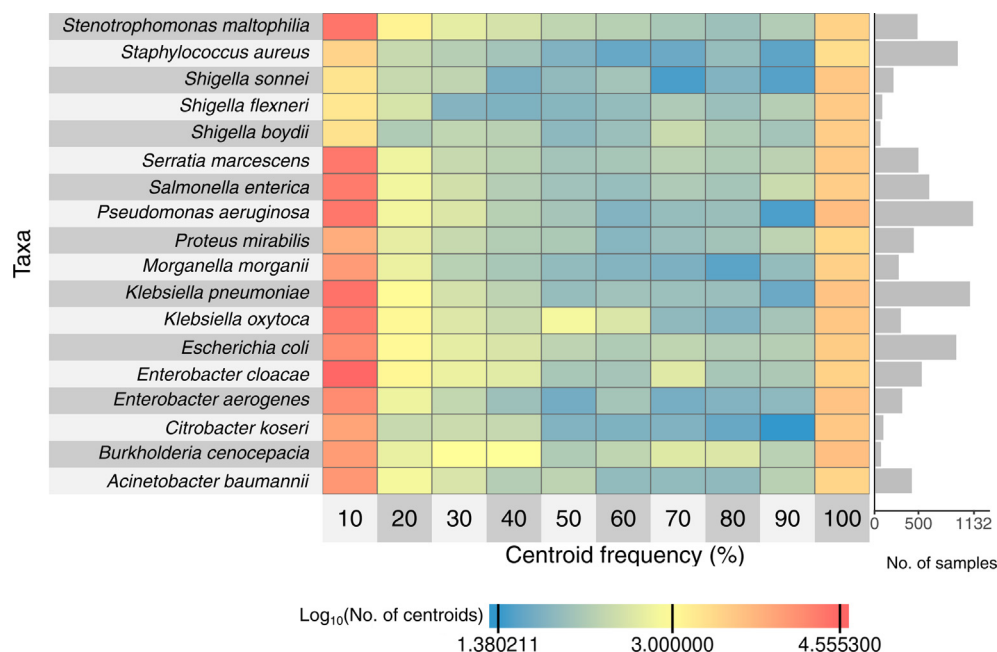


Figure 4 Centroid frequency

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 $\geq 90\%$ 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.

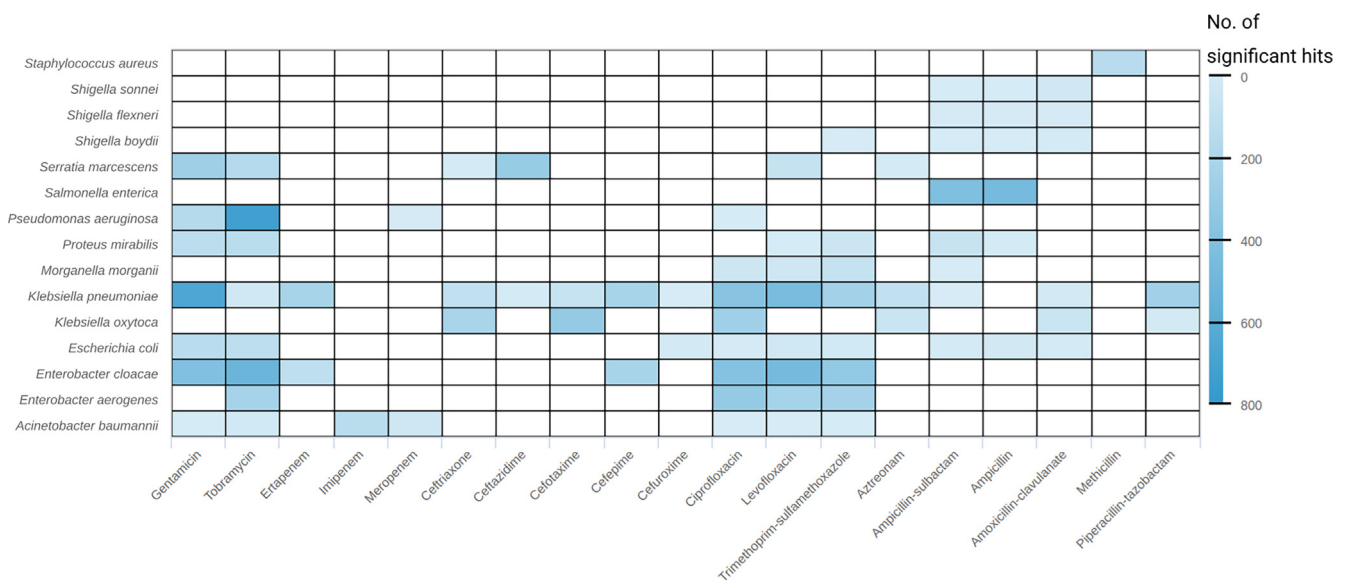


Figure 5 Number of significant results of the resistance association analysis

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 multi-disciplinary 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://hmmmer.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://hmmmer.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 ≤ 1000 , N50 ≥ 5000 , and L50 ≤ 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. morgani*, *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 -distancel kbit20_3). HMM profiles were calculated using hmmbuild from the HMMER software package (<http://hmmmer.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://hmmmer.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 ≥ 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 $\geq 99\%$ 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 $\geq 90\%$, $\geq 95\%$, and $\geq 99\%$ 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 $\geq 90\%$ 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 = $1 - \text{cor}^2$, 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 (λ) was < 1.1 for the smallest possible k . If none of the computed λ values was < 1.1 , then k with the smallest λ 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-dad72dddf5a/>). 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 constraints, 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, Weshnowski B, Vashisht R, Taylor 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, Massignani 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-adapted 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, Ciufu 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–e817.
- [41] Hennig G, Gehrman 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.