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 extrachromosomal 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.

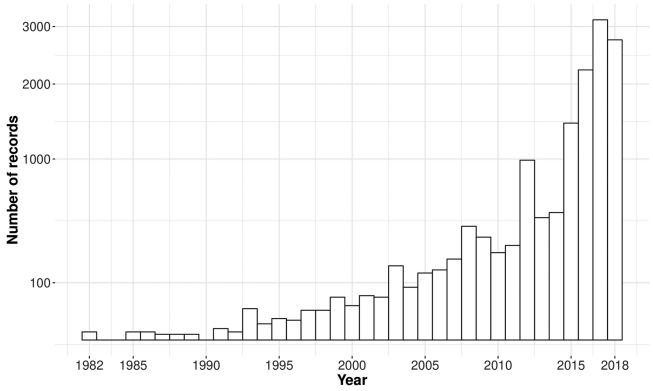

Figure 1. Number of plasmid records included into the collection grouped by the year of their creation. The *y*-axis scale is square root transformed.

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).

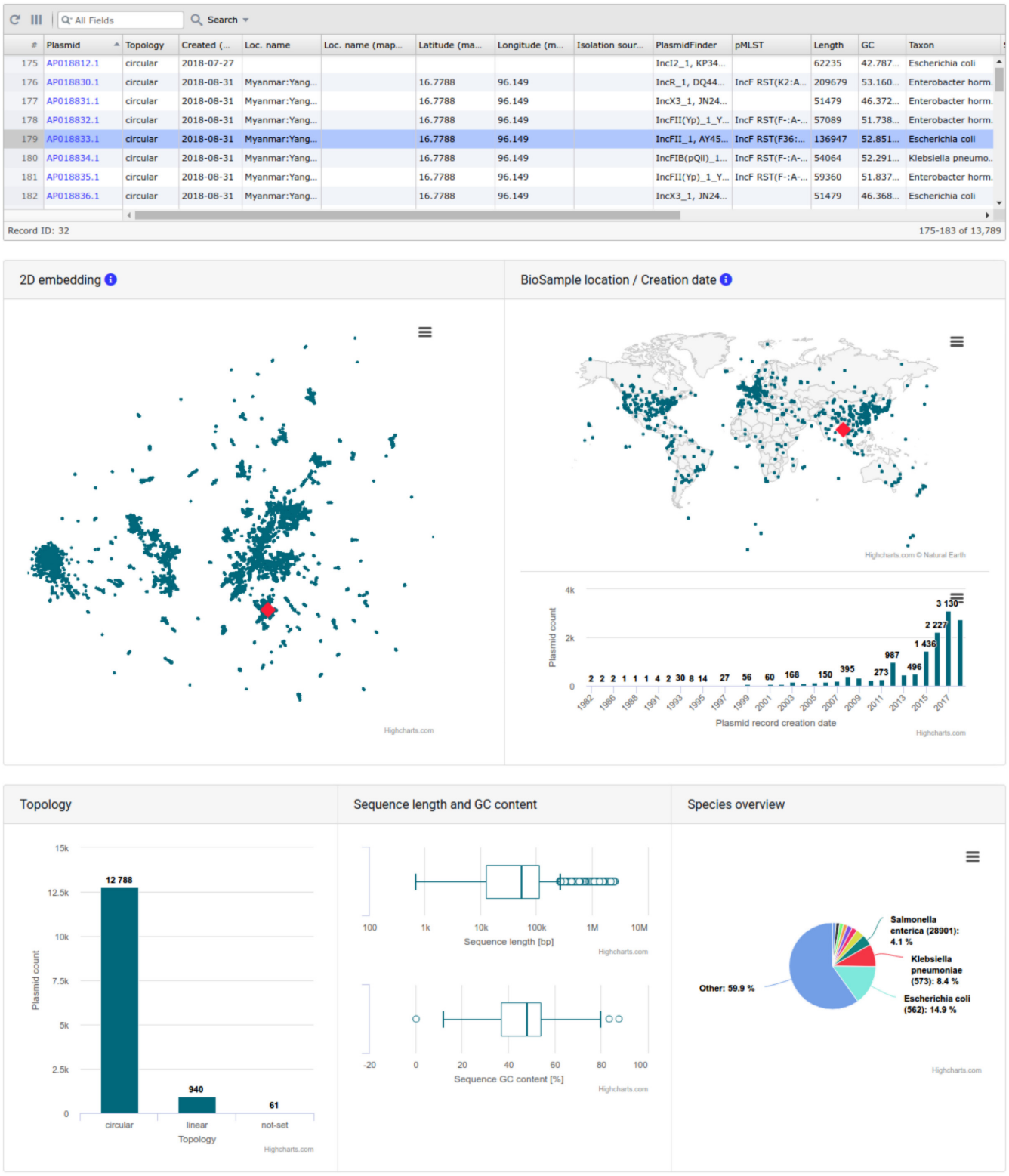

Figure 2. Interactive overview of collected plasmid records. Record AP018833.1 is selected in the table and highlighted (red diamond shaped symbol) in the embedding plot and on the world map.

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 asso-

ciated *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.