BacARscan: A Comprehensive and Interactive Web-Resource to Discern Antibiotic Resistance Gene Diversity in –Omics Datasets

Deeksha Pandey
University of Delhi South Campus

Bandana Kumari
University of Delhi South Campus

Neelja Singhal
University of Delhi South Campus

Manish Kumar (✉ manish@south.du.ac.in)
University of Delhi South Campus

Research Article

Keywords: Antibiotic resistance, Surveillance tool, Monitoring, Metagenome, Hidden Markov Models

DOI: https://doi.org/10.21203/rs.3.rs-712636/v1

License: ©  This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Regular surveillance of antibiotic resistance genes (ARGs) is important to understand the emergence and epidemiology of antibiotic resistance (AR) in clinical and environmental niches. With diminishing costs, NGS technologies are anticipated to replace classical microbiological and molecular methods for determination of AR. One major hindrance underlying identification and annotation of ARGs from WGS data is that a major part of genome databases contain fragmented genes/genomes (due to incomplete assembly). Herein, we propose a web resource of Bacterial ARGs, named as BacARscan (Bacterial Antibiotic Resistance scan), to detect, predict and characterize ARGs in metagenomic, genomic and proteomic data. The current version of BacARscan comprises 254 ARG models, each annotated with a resistant profile against different classes of antibiotics, resistance mechanism etc. Benchmarking on a combined dataset of AR and non-AR proteins found 92% precision & 95% F-measure. BacARscan can also discriminate between the protein families that are homologous but not all families are involved in the AR. BacARscan identified more ARGs in (a) gut microbiome and (b) datasets comprising short read genomic and proteomic sequences of ESKAPE pathogens. Analysis of clinical metagenomic data indicated its potential to complement and/or supplement WGS based identification of ARGs in clinical samples. BacARscan standalone software and web-server are freely available at http://proteininformatics.org/mkumar/bacarscan and github repository (https://github.com/University-of-Delhi-south-campus/BacARscan).

Introduction

Antibiotics, often designated as ‘miracle drugs’, revolutionized the prevention and treatment of bacterial infections\(^1\). However, irrational and indiscriminate use of antibiotics has resulted in evolution of antimicrobial resistance against almost all major classes of antibiotics\(^2,3\). Antibiotic resistance is not restricted to any particular geographical region or country, rather it is a global phenomenon, which over the last several years has spiraled into a major problem of public health significance. Realizing the crisis created by expansion of microbial antibiotic resistance and non-significant antibiotic development efforts, the World Health Organization in 2017, published the “first global priority pathogen list” and categorized antibiotic-resistant bacteria into critical, high and medium priority pathogens, according to the urgency of need for new antibiotics\(^4\).

Antibiotic resistance is a dynamic phenomenon. Hence, regular surveillance of antibiotic resistance genes of microbes and metagenomes, spanning clinical and environmental settings is important to understand the epidemiology and foresee the emergence of new antibiotic resistance factors. The rapid decline in the cost of high-throughput DNA sequencing technologies has resulted in their use in routine clinical\(^5\)\(^\text{–}\)\(^7\) and environmental samples. Consequently, a plethora of information about antibiotic resistance and antibiotic resistance genes (ARGs) is available in gene databases. Also, metagenomic studies have added immensely to our knowledge of ARG pool\(^8\). It is expected that in the near future, classical microbiological and molecular methods for determination of antimicrobial resistance might be replaced by whole genome sequencing (WGS) based identification of microbial resistomes using antibiotic resistance databases and \textit{in-silico} prediction tools. However, at present a major hindrance in identification and annotation of ARGs from WGS data is the abundance of fragmented genes/genomes in the genome databases (due to incomplete assembly). This also
implies that an ideal bioinformatics tool, which can detect ARG in whole genomes, should be able to work effectively on short sequencing reads and fragmented contigs, too.

In the past few years, many in-silico resources and tools have been developed to predict, monitor, catalogue and characterize the spread of ARG\(^1\). This includes Antibiotic Resistance Genes Database, ARDB\(^9\), Comprehensive \(\beta\)-lactamase Molecular Annotation Resource, CBMAR\(^10\), ResFinder\(^11\), Comprehensive Antibiotic Resistance Database, CARD\(^12\), and Resfams\(^13\). These tools have provided important insights in identification and prevalence of ARG and prediction of new resistance mechanisms. Among different antibiotic resistance databases, CARD is the most popular database among microbiologists owing to its free public availability and ease of use.

It has been reported by several people that the existing bioinformatics tools of ARG prediction/identification are biased towards a few specific ARGs. For example, ResFinder\(^11\) and SEAR\(^14\) are more tuned to predict the plasmid-based ARGs\(^15\). Another popular ARG resource, PATRIC, is suitable to identify resistance against carbapenem, methicillin, and \(\beta\)-lactam antibiotics.

Arango-Argoty et al. has shown that a similarity search using the manually curated potential ARGs from UniProtKB database against the CARD and ARDB genes resulted in sequence identity in range of 20 to 60\% terms of the percentage identity, bit score >50, and highly significant e-value (< 1e-20)\(^16\). This indicated that ARG annotation tools based on the “best hit” approach might suffer from a high rate of false negatives. In general a high identity cut-off (>80\%) is used as a threshold to determine the diversity of ARGs in a resistome. This means that a significant number of ARGs might be declared as false negative during similarity search.

The problem of false negatives is further compounded in case of short reads consisting of nearly 100bp or 25 amino acids. To avoid a high false positive rate generally a very high identity threshold (~80\%) is considered\(^17\), which might result in a high rate of false negative ARG annotation.

It has been also observed that very stringent criteria of similarity search based ARG annotations would mostly identify known ARGs\(^11,13\). This means several new ARGs that were discovered by functional screening, would have been missed by the similarity-based approach\(^18–23\). The results also indicate that the best-hit similarity search approach (a) is more suitable to find only highly conserved ARGs, and (b) might fail to find novel and/or those with low sequence identity to known ARGs.

The abundance of ARGs in metagenomic and transcriptomic data is done by alignment of individual reads to a reference database or by de-novo assembly of short-reads in large contigs and then alignment to a reference database. Though alignment after short-read assembly is considered as a better method of annotation, in case of low-coverage of reads, the assembly is highly fragmented hence similarity search based methods are not very useful. Due to the ability to efficiently model the conservation of a sequence, Hidden Markov models (HMMs) can identify the regions that are necessary for protein function. Hence a HMM based tool might be more sensitive in detection of remote homology. Two most popular databases for protein annotation, namely, Pfam and TIGERFAM, are examples of the better annotation capability of HMM in comparison to simple sequence alignment tools. However, very limited application of profile HMMs has been done in ARGs annotation. One such example of usage of HMM in ARG annotation is Resfams\(^13\), which is a HMMs based search tools. Resfams offer an alternative to the pairwise local sequence alignment based tools for
identification of emerging and/or unknown ARG on the basis of homology to functional domains of AMR protein. It has been reported that HMM based method, namely Resfams, identified AMR proteins with a low false-discovery rate vis-à-vis BLAST based approach of AMR protein annotation using ARDB and CARD\textsuperscript{13}.

Despite several advantages of HMMs over the simple alignment based search tools, a very limited attempt has been made to apply profile HMMs in ARGs annotation. Here, we have described an \textit{in-silico} tool, named as BacARscan (acronym for “Bacterial Antibiotic Resistance scan”), for quick identification and annotation of ARGs in –omics datasets. BacARscan is developed using modeling of insertions, deletions and conservation of amino acids in the form of profile-HMM. Each profile-HMM of BacARscan is annotated with relevant details such as the class of antibiotics targeted by the gene and functional type of ARG genes using the existing ARG databases. It is user friendly, freely accessible and can be integrated in any user defined ARG annotation pipeline. BacARscan is composed of two types of ARG HMM libraries namely pARGhmm and nARGhmm that can work on protein and gene dataset respectively. During the search, BacARscan scans all profile-HMMs against the query sequence and lists the ARGs in the query dataset and its corresponding details such as the class of antibiotics targeted by the gene and functional type of ARG genes.

Materials And Methods

Retrieval of antibiotic resistance gene and protein sequences

The data on antibiotic resistance genes/proteins was collected from various databases that contains information about antibiotic resistance genes/proteins such as ARDB\textsuperscript{9}, ARG-ANNOT \textsuperscript{24}, CARD\textsuperscript{12}, CBMAR\textsuperscript{10}, INTEGRALL\textsuperscript{25}, RAC\textsuperscript{26}, Tetracycline + MLS nomenclature, UCARE\textsuperscript{27}, Lahey Clinic, Resfams\textsuperscript{13}, ResFinder\textsuperscript{11}, HMP\textsuperscript{28}, LacED\textsuperscript{29}, MvirDB\textsuperscript{30}, Institut Pasteur and Patric\textsuperscript{31}. Some of these databases are focused only on the ARGs while few provide the ARG information as a part of the larger schema such as integrons (INTEGRALL), Antibiotic resistance Cassettes (RAC). In addition to the antibiotic resistance databases, we also collected data from the published research papers. All ARGs were first binned on the basis of their antibiotic inactivation profile. Members of each bin were then divided into distinct clusters on the basis of sequence homology using the BLASTClust module of the NCBI’s BLAST program at identity threshold of 90% and query coverage of 95% (BLASTCLUST with “-S 90 and -L 0.95” option) to cluster the protein sequences. After removing protein clusters that were composed of only fragmented sequences, we got 254 sequence clusters.

Building of pARGhmm profile from antibiotic resistance protein sequences

Firstly we did cluster-wise global multiple sequence alignment (MSA) of proteins using MUSCLE\textsuperscript{32} at default parameters. Each MSA was manually curated to remove the fragmented sequences. Each MSA was then converted into profile-HMM using the hmmbuild function of HMMER (version 3.1) at default settings\textsuperscript{33}.

The library of profile-HMM models of protein sequences of ARGs was built here henceforth referred as pARGhmm. To assess the ability of HMM models to accurately identify the proteins of its own family we used a dataset composed of ARG and non-ARG sequences. We observed that with an e-value threshold 1e-6 we were able to assign the maximum number of proteins to the correct ARGhmm. Hence for all subsequent
searches we used 1e-6 as the e-value threshold. When multiple ARGhmm were aligned to the same region of a protein with e-value lower than 1e-6, the model with the maximum alignment score was chosen as the best ARG model.

**Functional annotation of pARGhmm profiles**

All HMM models of pARGhmm were annotated with their functional properties using the databases from which ARG sequences were obtained and the published literature. The information that was associated with each ARGhmm are (a) class and subclass of antibiotics against which the query proteins/genes impart resistance, (b) resistance mechanism, (c) antimicrobial resistance spectrum, (d) AMR protein name & families (e) function of AMR genes. Besides the UniProt ID of the proteins from which annotations were collected, is also mentioned (Table S1).

**Construction and functional annotation of nucleotide version of ARGhmm (nARGhmm)**

To facilitate usage of ARGhmm directly on genomics/metagenomics data, without translating them into the protein sequences, we also developed a nucleotide version of ARGhmm. It was named as nARGhmm. The procedure to build nARGhmm was the same as what we followed during building of pARGhmm. Briefly, for each AR sequence cluster, we constructed two different HMMs, one for protein (pARGhmm) and another for nucleotide (nARGhmm). The complete set of 254 functionally annotated n and pARGhmm were named as BacARscan. The schematic schema of building (pARGhmm & nARGhmm) from a dataset of curated antibiotic resistance protein/gene sequences is shown in (Fig. 1).

The performance of nARGhmm was evaluated on a dataset of short sequence reads generated from the back-translation of protein sequences of positive and negative independent dataset. The length of each read was 100 nucleotides and from each ARG sequence 20 random short sequence reads were generated. Similar to the pARGhmm, the prediction capability of nARGhmm was also evaluated using precision and F-score.

**Benchmark datasets**

**Dataset-I: Evaluation dataset**

The efficiency of BacARscan was evaluated using a dataset consisting of two different categories of sequences namely, positive and negative. The negative sequences were obtained from protein clusters formed by BLASTclust and had < 5 sequences. Since, these proteins were not used to create the profile-HMM models hence, they were designated as negative datasets. Sequences that were used to build the pARGhmm i.e. protein clusters which contained ≥ 5 sequences were called as positive dataset.

**Dataset-II: Independent dataset**

A. To test the performance of BacARscan vis-a-vis other methods of ARG annotation, we also did a comparative performance evaluation on an independent dataset obtained from NCBI Bioproject. This dataset contained gut microbiomes of 13 healthy individuals of various ages and genders, including an
unweaned infant (BioProject Accession: PRJNA28117) (Taxonomy ID 408170)\(^ {34}\). The dataset has a total 1,67,118 gene, which were translated to the protein.

B. For the benchmarking purpose, we used two independent datasets for evaluating the prediction capability of BacARscan. 1) Penicillin-binding proteins (PBPs) or DD-peptidase proteins, which were retrieved from UniProtKB database dated on 15-02-21 using keyword search (penicillin binding proteins). We found 60 reviewed non-fragmented PBPs in the UniProtKB database whose existence was established at protein level. 2) Non-antibiotics resistant bacterial efflux (non-ARE) proteins that were collected from our earlier work (Reference). The final benchmark independent datasets consisted of 60 PBPs and 389 Non-ARE protein sequences respectively.

**Dataset-III: Annotation of ARGs in different strains of ESKAPE pathogen**

The ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* species) are globally recognized as leading cause of nosocomial multidrug-resistant bacterial infections. To assess the ability of BacARscan to discern the diversity of ARGs among the members of different ESKAPE pathogens, and strains therein, we annotated five strains of every ESKAPE pathogen using BacARscan. For annotation we used proteomes of those ESKAPE pathogens whose genome is completely sequenced and the complete chromosome and plasmid protein sequences were available at NCBI. The detail of various strains of ESKAPE pathogens used for evaluation is mentioned in Table S2.

**Dataset IV: Annotation of clinical metagenomic data**

To demonstrate the usefulness of BacARscan in understanding the resistome composition of a clinical sample, we used a metagenomic dataset that was obtained from six human patients who undertook cholecystectomy for cure of acute cholecystitis\(^ {35}\). The short-read sequences of these studies were obtained from DNA Databank of Japan (Accession: DRA005134). The original study\(^ {35}\) was conducted to estimate the bacterial infection in inflamed gallbladder. The study cohort consisted of patients of various age groups (range 43–85 years, median age 64.5 years) and an equal number of male and female patients. From each patient sample, metagenomic study was performed on bile, feces and saliva samples. It was reported in the original study that from one patient, saliva and feces samples were not obtained, which resulted in 16 samples instead of 18. In total there were 16 metagenomic samples, six from human bile and five from gut and saliva each.

**Evaluation parameters**

During evaluation only hits with e-values lower than 1e-6 were considered as significant. The e-value cut-off of 1e-6 was determined on the basis of independent evaluation of ARGhmm. At e-value 1e-6 we found maximum precision, recall and F-score (details in Case Study-I: Comparative evaluation of prediction efficiency of pARGhmm). To evaluate the efficiency of both n & pARGhmm, we classified the prediction results in four different categories namely, true positives and true negatives (TP and TN respectively), false positives and false negatives (FP and FN respectively). TP and TN were the observations that were correctly predicted, while the predictions categorized in FP and FN were wrong. When an ARG sequence was predicted to the actual
class to which it belongs, the prediction was classified as TP. Similarly, when a non-ARG was predicted as non-ARG, it was classified as TN. When an ARG and non-ARG were predicted as non-ARG and ARG respectively, the prediction result was labeled as FN and FP, respectively. Using the above-described classification we calculated precision, recall and F-score to measure the prediction capability of BacARscan.

i. Precision: It is the ratio of correctly predicted positive observations to the total number of predicted positive observations. High precision indicates a low false positive prediction. Precision was calculated as

\[
\text{Precision} = \frac{TP}{TP + FP} \times 100
\]  

ii. Recall or Sensitivity: It is the ratio of correctly predicted positive observations to the total number of positive samples submitted for prediction. High recall indicates a highly efficient predictor with a very low number of false predictions. It was calculated as:

\[
\text{Recall} = \frac{TP}{TP + FN} \times 100
\]  

iii. F-measure or F₁ or F-score: In order to balance the precision and recall values due to unequal composition of positive and negative datasets, F-score was also calculated. It is basically harmonic mean of precision and recall and calculated as:

\[
F = 2 \times \frac{(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}
\]

Web interface and standalone tool

The web version of BacARscan is available without any restriction and can be accessed at http://proteininformatics.org/mkumar/bacarscan. A schematic diagram to explain the working of BacARscan is presented as Fig. 8. In the BacARscan web tool, a user has the option to choose between query sequence type and nature of HMM-profiles (either ‘Protein/pARGhmm’ or ‘Gene/nARGhmm’). The web platform of BacARscan can process only 10 sequences at a time and the processing takes only a few seconds. To annotate a large dataset standalone version of BacARscan is also available that can be downloaded from the download section of BacARscan webtool. We also quantified the annotation speed of the standalone version of BacARscan on a dataset, which had five different strains of all six ESAKPE pathogens. This means a total 30 complete proteomes (6 organisms × 5 different strains) containing 1,28,305 protein sequences were used during speed assessment. It took nearly 31 minutes to complete the annotation of all 30 proteomes, which means the average time of annotation was one minute per proteome. All speed assessment experiments were run on a computer with an Intel(R) Xeon(R) 4 Core E5507 2.27 GHz processor with 6 GB DDR4 RAM, running 64-bit Red Hat Enterprise Linux operating system (Release 6.2). The results of speed assessment clearly indicate that BacARscan is a powerful, fast and cost-effective tool of annotation of ARG in WGS data compared to other in-silico tools.

Results & Discussion
Performance assessment of pARGhmm

To assess performance of pARGhmm, we used an evaluation dataset that consisted of antibiotic resistant (positive dataset) and non-antibiotic resistant proteins (negative dataset). When the first hit of pARGhmm search result on the basis of e-value, was used to evaluate the prediction capability, 228 true positives and 26 false positives predictions were found (Table 1). We also evaluated the performance of pARGhmm using a majority approach by including multiple search results from the top and decided the final outcome by the consensus approach. For example, during assessment on the basis of top three search results, the sequences which are in the top three hits and belong to the class of query pARGhmm, they were categorized as true positive and the sequences which did not belong to the class of query pARGhmm, was categorized as false positive. Since the majority of search results (two out of three) belong to the class of query pARGhmm hence overall the prediction would be categorized as true positive. With the majority approach the maximum performance was achieved when top five hits were used to evaluate the performance. The precision and F-measure was 92.12% and 95.90% respectively. Overall the assessment results showed that the prediction performance increased till the decision was made on the basis of five hits, afterwards the performance started decreasing. The consistency in search efficiency of each ARGhmm was evaluated using the leave-one-out cross validation (LOOCV) approach. During LOOCV one sequence was excluded from the process of model building and then the performance of model was tested against the dataset containing the (a) excluded sequence, (b) sequences using which remaining 253 HMMs were build and (c) non-antibiotic resistance sequences that were not used for making of ARGhmm. We found near perfect consistency in performance of the majority of HMMs (Figure S1).
Table 1
Performance of BacARscan (pARGhmm & nARGhmm)

| No. of top hits | pARGhmm | nARGhmm |
|-----------------|---------|---------|
|                 | True Positive | False Positive | Precision | F-measure | True Positive | False Positive | Precision | F-measure |
| 1               | 228      | 26       | 89.76    | 94.60     | 231         | 23       | 90.94     | 95.25     |
| 3               | 229      | 25       | 90.15    | 94.82     | 235         | 19       | 92.51     | 96.11     |
| 5               | 234      | 20       | 92.12    | 95.90     | 237         | 17       | 93.30     | 96.53     |
| 7               | 233      | 21       | 91.73    | 95.68     | 236         | 18       | 92.91     | 96.32     |
| 9               | 232      | 22       | 91.33    | 95.47     | 240         | 14       | 94.48     | 97.16     |
| 11              | 209      | 45       | 82.28    | 90.28     | 241         | 13       | 94.88     | 97.37     |
| 13              | 182      | 72       | 71.65    | 83.48     | 240         | 14       | 94.48     | 97.16     |
| 15              | 158      | 96       | 62.20    | 76.69     | 238         | 16       | 93.70     | 96.74     |

The overall assessment of BacARscan HMMs profile: Scanning of AR genes through profiles HMMs against protein sequences (positive and negative dataset) and gene sequences (reads) i.e. pARGhmm & nARGhmm. First column showing couple of alternative tophits like top1st, 3rd, 5th, 7th hits and so on. True Positive & False Positive is the calculation of accuracy of our prediction method. Precision known as exactness which is calculated with above values of true & false positive rates respectively; F measure known as harmonic mean, calculated with the help of above scores.

Comparative evaluation of prediction efficiency of nucleotide and protein modules of BacARscan (p & nARGhmm)

The performance of nARGhmm was evaluated using a dataset of short read sequences of 100 nucleotide length. The short read sequence library was generated from gene sequences constituting the positive and negative evaluation datasets. When only the first hit was considered for evaluation, the precision and F-measure values were 90.94 and 95.25 respectively (Table 1). Similar to pARGhmm, the performance of nARGhmm was also evaluated using the majority approach. The best performance with majority approach was achieved when top five hits were used to quantify the performance. The comparative evaluation of performance of pARGhmm and nARGhmm also indicate comparable performance in terms of both precision and F-measure. Based on the above findings, we can conclude that BacARscan is an efficient tool to find ARGs in both protein and gene datasets.

Case Study-I: Comparative evaluation of prediction efficiency of pARGhmm on metagenomic data

The prediction efficiency of pARGhmm vis-à-vis other annotation tools was evaluated on an independent dataset obtained from a functional metagenomic study obtained from the gut microbiome of 13 healthy humans\textsuperscript{34}. BacARscan identified 235 types of ARG in it (Table 2). Other methods namely, BLAST search
against ARDB found 155 types of ARG (Table 2), Resfams revealed 166 different types of ARG, while Resistance Gene Identifier (RGI) tool of CARD found 157 types of ARGs.

Table 2

| Dataset             | Nucleotide sequences | Translated sequences | Programs | Predicted types of AR genes |
|---------------------|----------------------|----------------------|----------|-----------------------------|
| Gut Metagenome      | 1,67,118             | 10,02,708            | ARDB     | 155                         |
|                     |                      |                      | CARD     | 157                         |
|                     |                      |                      | Resfams  | 166                         |
|                     |                      |                      | BacARscan | 235                        |

Comparative analysis of prediction of different types of AR genes: Dataset composed with nucleotide and translated sequences of Gut metagenome, multiple programs (ARDB, CARD, Resfams and BacARscan) searched against gut metagenomic dataset. Different types of AR genes predicted using diverse methods respectively.

Comparison on the basis of resistance mechanism showed that in four mechanism classes namely acetyltransferase, efflux, beta-lactamase and antibiotic target alteration the number of predicted ARGs were more than Resfams. In ‘gene modulating resistance’ and ‘glycopeptide resistance’ the number of distinct profiles was more in Resfams. The results showed that the performance of BacARscan was better than Resfams, which is also a HMM based ARG annotation tool. Moreover, in BacARscan antibiotic-resistance-mechanism categories have been further sub-classified so that it can deliver the maximum available information to the user (Figs. 2 and 3). Our results indicated that not only was the performance of BacARscan was better; it provided more information than other ARG annotation tools.

**Case Study-II: Annotation of antibiotic resistance genes in ESKAPE pathogens**

The results of ARG annotation in all five ESKAPE pathogens revealed predominance of efflux based ARG, followed by beta-lactamases or acetyltransferases based ARG. The diversity of ARG in ESKAPE pathogens discerned by BacARscan was similar to other similar work. Though a slight variation was found in the number of ARG among different strains of each ESKAPE pathogens, however the distribution of different ARG classes was similar.

We also reassessed the resistome prediction ability of other ARG annotation tools vis-à-vis BacARscan using the ESKAPE pathogen data. Resfams annotations revealed that the majority of ARGs present in the ESKAPE pathogens belonged to the efflux category (Fig. 4). Although both Resfams and BacARscan discerned a similar pattern of distribution of ARG, BacARscan found a significantly greater diversity in ARG encoding efflux proteins than Resfams. On the basis of homology to CARD database proteins, the output of RGI-CARD was divided in three different categories namely, perfect, strict and loose. In the present work only ARG of perfect and strict category were included for comparison. Similar to BacARscan and Resfams, as per RGI-CARD the majority of ARG belonged to the efflux category. But, the number of ARG predicted by RGI-CARD and ARDB was very less in comparison to BacARscan and Resfams (Table S4).
Case Study-III: Comparison of BacARscan and UniProtKB ARG annotations

To validate the ARG annotations done by BacARscan, we used *Acinetobacter baumannii* strain 6200 (Ab6200) [Genome Assembly ID: 216230] as a model. We compared the ARG annotations of Ab6200 in UniProtKB with BacARscan annotations. Our results showed that BacARscan annotated many proteins of Ab6200 related to antibiotic resistance, which were absent from the list of antibiotic resistance proteins annotated by UniProtKB (Table S3). This indicated that BacARscan is capable of identifying more and unannotated antimicrobial resistance factors in microbial pathogens, however the novel ARGs predicted by BacARscan needs experimental verification.

Case Study-IV: Prediction of antibiotic resistance genes in clinical genomic samples

We performed ARG search against short read sequences of metagenomic samples that were obtained from bile, gut and saliva samples of six human patients (named from P1 to P6) suffering from acute cholecystitis (25) using nARGhmm. The original metagenomic analysis was performed by Kujiraoka et al.\(^\text{35}\) to discern the imbalance in the microbiota of patients suffering from acute cholecystitis. As shown in Fig. 5, we found a large number of ARGs in all three samples of patient P1. Interestingly we didn't find any ARG in patient P5 and only two ARGs in P6. Overall we found the number of ARGs was highest in the gut microbiome followed by bile and saliva microbiomes. We feel this may be due to high microbial load in the gut sample as observed in the original study also. We also noticed that efflux was the most abundant class of ARG across the microbiomes. The fact that BacARscan did not identify ARG in the bile microbiomes of patient P5 and P6 indicates it’s accuracy of prediction and potential to complement and/or supplement WGS based identification of ARG in clinical samples. This observation was inline with the original study that reported very less number of bacterial reads in bile microbiome of patients P5 and P6\(^\text{35}\). The results showed that BacARscan could discern the ARG diversity and the corresponding mechanism of antibiotic resistance in clinical metagenomic studies also. We could not compare our results with any other ARG annotation methods, since none of the existing tools provide a platform to scan the sequences containing raw non-assembled reads.

Comparison with other existing ARG annotation resources

Among several ARG annotation resources, ARDB, CARD and Resfams are the most popular ones. Hence, we also compared the performance of BacARscan *vis-à-vis* ARDB, CARD and Resfams. ARDB\(^\text{9}\) was the first *in-silico* resource that provided a centralized repository to facilitate characterization of ARGs. It kick-started the annotation of ARG sequence data obtained from high-throughput sequencing using BLAST. The Comprehensive Antibiotic Resistance Database (CARD) is the most updated repository of ARG\(^\text{12, 38}\). It identifies and annotates ARG either using BLAST or Resistance Gene Identifier (RGI) module. The BLAST option performs sequence similarity search against the CARD reference sequences. RGI uses the Resistance Gene Identifier to predict ARG(s) based on homology and SNP models. Resfams is a database of protein families, which exhibit antibiotic resistance\(^\text{13}\). It tried to improve over the simple sequence alignment based
methods by adding position specific information to make local alignments. In Resfams, each ARG is represented as HMM models.

ARDB and CARD have an integrated BLAST program to infer the biological roles of proteins, based on their similarity to experimentally characterized proteins. Simple linear sequence alignment based methods are very effective for comparing sequences with a high degree of similarity (60 percent or higher) but fail in case of highly divergent and distantly related proteins. Also, the last update of ARDB was done in 2009, which presently makes it less attractive than other highly curated AR databases like CARD or Resfams. In BacARscan improvements were made by (a) adding position specific information to each alignment, (b) measuring sequence similarity against curated domain assignments and (c) including probabilistic models namely, profile Hidden Markov Models which calculates a sequence profile on the basis of frequency of occurrence of a given amino acid at each position of a protein sequence. Both Resfams and BacARscan have used profile based searching, which has a better sensitivity and specificity, faster and more accurate, and a better ability to detect highly diverged homologs due to the strength of its underlying probability models in comparison to BLAST. Hence, BLAST-based search is more suitable, quick and easy for smaller datasets, while BacARscan is inherently more apt to handle big datasets generated by metagenomic or transcriptomics study.

Though, both BacARscan and Resfams are based on HMM models, BacARscan has six advantages over Resfams (a) BacARscan is available both as web server and standalone tool; (b) the number of profile-HMM models in BacARscan is 254, which is nearly 50% more than 170 profile-HMM models in Resfams; (c) BacARscan can be used on both protein and nucleotide sequence data with comparable performance efficiency; (d) The nucleotide version of BacARscan can be used on a fully assembled genome as well as short read library without any need for assembly of reads; (e) BacARscan has significantly more ARGhmmms representing each antimicrobial resistance mechanism than Resfams (Fig. 6). Further we have also categorized the BacARscan ARGhmmms on the basis of their target antibiotics (Fig. 7).

**Benchmarking on homologous non-antibiotic resistance conferring proteins**

In order to assess the capability of BacARscan to predict proteins/sub-families that doesn't provide antibiotic resistance, but homologous proteins/sub-families can have antibiotic resistance capability, we constructed an independent dataset comprised of Penicillin-binding proteins (PBP s) and non-antibiotic efflux proteins (non-ARE). Both PBP and β-lactamases belong to the superfamily of serine penicillin-recognizing enzymes and have similar conserved protein folds. It is pertinent to mention here that no PBP s and Non-ARE proteins were used to construct the HMM models. Similar to other previous studies, during the benchmarking we have adopted the best-hit approach with an e-value cutoff 1e-20. When we scanned this database using BacARscan, six PBP s were annotated as β-lactamases while 23 non-ARE proteins were annotated as efflux based ARGs. Overall the results showed that BacARscan could discriminate between the ARG and non-ARG homologous proteins/sub-families.

**Description of the BacARscan webserver**

To provide the BacARscan prediction module to the scientific community, we also established a webserver at which gene/protein sequences can be uploaded for annotation. The output of BacARscan contains the
complete annotation of the query sequence. A maximum of ten sequences can be processed by BacARscan webserver in one go. For whole genome/metagenomic/proteome scale annotation, standalone version will be required or provided in the download section of the webserver or from Github repository.

**Potential use of BacARscan**

Most traditional antibiotic resistance determination methods are based on laboratory tests but the time and resources required to generate this data sometimes make it unsuitable for quick monitoring. The development of high throughput sequencing methods has opened an alternative way of quick and cost-effective method of antibiotic resistance determination. BacARscan is an effort to provide an additional way forward to identify the ARGs in an -omics (proteomics/genomics and metagenomic) sample. BacARscan can also be combined with the traditional surveillance and thus can complement the traditional methods of ARG annotation. The current version of BacARscan supports prediction using only 254 ARG families but in future, we will extend it to the new ARGs or families. We hope that BacARscan will help in the prediction of ARGs and it will help in progress of studies related to AMR.

**Conclusions & Further Developments**

In the present study, we have described a novel antibiotic resistance analysis tool named BacARscan. The proposed tool can be used to monitor and annotate microbial ARG in both genomics and proteomic datasets. Benchmarking with other widely used online ARG annotation tools like ARDB and Resfams showed that BacARscan was able to identify more ARGs than these tools. One of the most notable improvements of BacARscan over other ARG annotation methods is its ability to work on genomes and small reads sequence libraries with equal efficiency and without any requirement for assembly of short reads. The analysis of ESKAPE pathogen proteomes showed that BacARscan could predict ARGs and provides valuable information about the mechanism of antibiotic resistance, thereof. Analysis of clinical metagenomic data set by BacARscan suggests its potential as a complementary and/or supplementary adduct for the traditional method of culture dependent analysis of ARGs. Benchmarking with non-AR proteins revealed that BacARscan has the ability to differentiate between ARG and non-ARGs. We anticipate BacARscan will be a valuable tool for the monitoring, characterization, identification and surveillance of ARGs in bacterial communities at an early stage of infection/outbreak. Also, BacARscan might be helpful for biologists in estimation and annotation of abundance of resistance gene(s) and their diversity in agricultural, environmental and animal-associated metagenomic samples and microbial communities. We believe that BacARscan would help the scientific community in their efforts to control the spread of antibiotic resistant bacteria in medicinal and clinical environments. Though the performance of BacARscan on different datasets looks promising, we understand that without a constant update, it would not remain useful to the scientific community. The current version of BacARscan supports prediction using only 254 ARG families but in future, we will extend it to the new ARGs or families. The information content of BacARscan will be periodically updated, as a new class of antibiotic resistance proteins will emerge. It should also be noted that none of the above-mentioned methods involve screening of AMR conferred via mutation but instead focus solely on dedicated resistance genes.

**Declarations**
Authors’ Contributions

DP and BK collected and organized the data and developed the web interface. DP, BK, NS and MK analyzed the results. MK conceived the idea and did overall supervision of the work. All authors reviewed the manuscript.

Competing Interest

The authors declare no competing interests.

Acknowledgements

All authors thank the University of Delhi (South Campus), New Delhi (India) for providing excellent facilities to carry out the research work.

Funding support

The work was carried out using the resources funded by the Science and Engineering Research Board under Fast Track Proposals for Young Scientists Scheme [Grant No: SR/FT/LS- 84/2010], UGC Major Research Project [Grant No: 41- 38/2012(SR)] and ICMR projects [ISRM/12(33)/2019 and VIR (25)/2019/ECD-1]. DP is supported by the Department of Science and Technology, Government of India under the INSPIRE Program [Grant Number: DST/INSPIRE 03/2015/003022]. NS was supported by the Council of Scientific Research under the Pool Scientist Scheme [Grant Number: 13(9089-A)/2019- POOL]. BK was supported by Indian Council of Medical Research under the Senior Research Fellowship Scheme [Grant Number: ICMR-BIC/11(33)/2014].

Data Availability

The tool and its dataset are freely accessible without any restriction at download page of web-server http://proteininformatics.org/mkumar/bacarscan/downloads.html

References

1. Xavier, B. B. et al. Consolidating and Exploring Antibiotic Resistance Gene Data Resources. J. Clin. Microbiol. 54, 851–859 (2016).
2. Baquero, F., Tedim, A. P. & Coque, T. M. Antibiotic resistance shaping multi-level population biology of bacteria. Front. Microbiol. 4, 15 (2013).
3. Chandra, H. et al. Antimicrobial Resistance and the Alternative Resources with Special Emphasis on Plant-Based Antimicrobials-A Review. Plants Basel Switz. 6, (2017).
4. Tillotson, G. A crucial list of pathogens. Lancet Infect. Dis. 18, 234–236 (2018).
5. Azarian, T. et al. Whole-genome sequencing for outbreak investigations of methicillin-resistant Staphylococcus aureus in the neonatal intensive care unit: time for routine practice? Infect. Control Hosp. Epidemiol. 36, 777–785 (2015).
6. Dominguez, S. R. \textit{et al.} Comparison of Whole-Genome Sequencing and Molecular-Epidemiological Techniques for Clostridium difficile Strain Typing. \textit{J. Pediatr. Infect. Dis. Soc.} \textbf{5}, 329–332 (2016).

7. Kinnevey, P. M. \textit{et al.} Enhanced Tracking of Nosocomial Transmission of Endemic Sequence Type 22 Methicillin-Resistant Staphylococcus aureus Type IV Isolates among Patients and Environmental Sites by Use of Whole-Genome Sequencing. \textit{J. Clin. Microbiol.} \textbf{54}, 445–448 (2016).

8. Schmieder, R. \& Edwards, R. Insights into antibiotic resistance through metagenomic approaches. \textit{Future Microbiol.} \textbf{7}, 73–89 (2012).

9. Liu, B. \& Pop, M. ARDB–Antibiotic Resistance Genes Database. \textit{Nucleic Acids Res.} \textbf{37}, D443-447 (2009).

10. Srivastava, A., Singhal, N., Goel, M., Virdi, J. S. \& Kumar, M. CBMAR: a comprehensive β-lactamase molecular annotation resource. \textit{Database J. Biol. Databases Curation} \textbf{2014}, bau111 (2014).

11. Zankari, E. \textit{et al.} Identification of acquired antimicrobial resistance genes. \textit{J. Antimicrob. Chemother.} \textbf{67}, 2640–2644 (2012).

12. McArthur, A. G. \textit{et al.} The comprehensive antibiotic resistance database. \textit{Antimicrob. Agents Chemother.} \textbf{57}, 3348–3357 (2013).

13. Gibson, M. K., Forsberg, K. J. \& Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. \textit{ISME J.} \textbf{9}, 207–216 (2015).

14. Search Engine for Antimicrobial Resistance: A Cloud Compatible Pipeline and Web Interface for Rapidly Detecting Antimicrobial Resistance Genes Directly from Sequence Data. \textit{PLOS ONE} \textbf{10}, e0133492 (2015).

15. McArthur, A. G. \& Tsang, K. K. Antimicrobial resistance surveillance in the genomic age. \textit{Ann. N. Y. Acad. Sci.} \textbf{1388}, 78–91 (2017).

16. Arango-Argoty, G. \textit{et al.} DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. \textit{Microbiome} \textbf{6}, 23 (2018).

17. Kleinheinz, K. A., Joensen, K. G. \& Larsen, M. V. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences. \textit{Bacteriophage} \textbf{4}, e27943 (2014).

18. Moore, A. M. \textit{et al.} Pediatric fecal microbiota harbor diverse and novel antibiotic resistance genes. \textit{PloS One} \textbf{8}, e78822 (2013).

19. Parsley, L. C. \textit{et al.} Identification of Diverse Antimicrobial Resistance Determinants Carried on Bacterial, Plasmid, or Viral Metagenomes from an Activated Sludge Microbial Assemblage. \textit{Appl. Environ. Microbiol.} \textbf{76}, 3753–3757 (2010).

20. Sommer, M. O. A., Dantas, G. \& Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. \textit{Science} \textbf{325}, 1128–1131 (2009).

21. Wichmann, F., Udikovic-Kolic, N., Andrew, S. \& Handelsman, J. Diverse Antibiotic Resistance Genes in Dairy Cow Manure. \textit{mBio} \textbf{5}, e01017-13.

22. Enault, F. \textit{et al.} Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. \textit{ISME J.} \textbf{11}, 237–247 (2017).

23. National Database of Antibiotic Resistant Organisms (NDARO) - Pathogen Detection - NCBI. https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/.
24. Gupta, S. K. et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob. Agents Chemother. 58, 212–220 (2014).
25. Moura, A. et al. INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. Bioinformatics 25, 1096–1098 (2009).
26. Tsafnat, G., Copty, J. & Partridge, S. R. RAC: Repository of Antibiotic resistance Cassettes. Database J. Biol. Databases Curation 2011, bar054 (2011).
27. Saha, S. B., Uttam, V. & Verma, V. u-CARE: user-friendly Comprehensive Antibiotic resistance Repository of Escherichia coli. J. Clin. Pathol. 68, 648–651 (2015).
28. Methé, B. A. et al. A framework for human microbiome research. Nature 486, 215–221 (2012).
29. Thai, Q. K., Bös, F. & Pleiss, J. The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. BMC Genomics 10, 390 (2009).
30. Zhou, C. E. et al. MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res. 35, D391-394 (2007).
31. Wattam, A. R. et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 42, D581-591 (2014).
32. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).
33. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37 (2011).
34. Kurokawa, K. et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes 14, 169–181 (2007).
35. Kujiraoka, M. et al. Comprehensive Diagnosis of Bacterial Infection Associated with Acute Cholecystitis Using Metagenomic Approach. Front. Microbiol. 8, 685 (2017).
36. [2010.16061] Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. https://arxiv.org/abs/2010.16061.
37. Brooks, L. E., Ul-Hasan, S., Chan, B. K. & Sistrom, M. J. Quantifying the Evolutionary Conservation of Genes Encoding Multidrug Efflux Pumps in the ESKAPE Pathogens To Identify Antimicrobial Drug Targets. mSystems 3, e00024-18 (2018).
38. Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 45, D566–D573 (2017).
39. Knox, J. R., Moews, P. C. & Frere, J. M. Molecular evolution of bacterial beta-lactam resistance. Chem. Biol. 3, 937–947 (1996).
40. Meroueh, S. O., Minasov, G., Lee, W., Shoichet, B. K. & Mobashery, S. Structural aspects for evolution of beta-lactamases from penicillin-binding proteins. J. Am. Chem. Soc. 125, 9612–9618 (2003).
41. Pandey, D., Kumari, B., Singhal, N. & Kumar, M. BacEffluxPred: A two-tier system to predict and categorize bacterial efflux mediated antibiotic resistance proteins. Sci. Rep. 10, 9287 (2020).

Figures
Figure 1

Steps adopted to construct profile-HMM models of BacARscan: An initial set of all AR proteins are curated and clustered. Compiled all clusters and used only those clusters which have ≥ 5 number of sequences in each cluster. Performed multiple sequence alignments of those clusters. MSAs are further used for creating profile HMMs of protein and nucleotides. Afterward each pARGhmm & nARGhmm annotated it via different sources. Statistics for each adopted step in the generation of the HMMs are in parentheses.
Figure 2

Number of ARGs and associated mechanisms found by ARDB, Resfams & BacARscan in human gut microbiome
Figure 3

Comparative analysis of CARD and BacARscan annotations of human gut microbiome on the basis of resistance mechanism
Figure 4

Comparison of prediction of ARGs and their resistance mechanism pattern between Resfams and BacARscan on various strains of bacterial communities (ESKAPE pathogens)
Figure 5
Comparison and prediction of ARGs on clinical genomic data

Figure 6
Comparison of Resfams and BacARscan profiles - HMM models on the basis of their resistance mechanism

Figure 7
Distribution of ARG HMM profiles into various antibiotic classes. The numerical value indicates the number of HMM that inactivates the antibiotic.

Figure 8
Schematic representation of BacARscan architecture

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.tiff
- TableS1.xlsx
- TableS2.docx
- TableS3.xlsx
- TableS4.xlsx