BacARscan: an in silico resource to discern diversity in antibiotic resistance genes

Deeksha Pandey,1 Bandana Kumari,1,2 Neelja Singhal1 and Manish Kumar1*

1Department of Biophysics, University of Delhi South Campus, New Delhi 110021, India
2Institute of Human Genetics-CNRS Montpellier, France

*Correspondence address. Department of Biophysics, University of Delhi South Campus, New Delhi, India 110021. Tel: +91-11-24157263; Fax: +91-11-24110876; E-mail: manish@south.du.ac.in

Abstract

Antibiotic resistance has escalated as a significant problem of broad public health significance. Regular surveillance of antibiotic resistance genes (ARGs) in microbes and metagenomes from human, animal and environmental sources is vital to understanding ARGs’ epidemiology and foreseeing the emergence of new antibiotic resistance determinants. Whole-genome sequencing (WGS)-based identification of the microbial ARGs using antibiotic resistance databases and in silico prediction tools can significantly expedite the monitoring and characterization of ARGs in various niches. The major hindrance to the annotation of ARGs from WGS data is that most genome databases contain fragmented genes/genomes (due to incomplete assembly). Herein, we describe an in silico Bacterial Antibiotic Resistance scan (BacARscan) (http://proteininformatics.org/mkumar/bacarscan/) that can detect, predict and characterize ARGs in -omics datasets, including short sequencing, reads, and fragmented contigs. Benchmarking on an independent non-redundant dataset revealed that the performance of BacARscan was better than other existing methods, with nearly 92% Precision and 95% F-measure on a combined dataset of ARG and non-ARG proteins. One of the most notable improvements of BacARscan over other ARG annotation methods is its ability to work on genomes and short-reads sequence libraries with equal efficiency and without any requirement for assembly of short reads. Thus, BacARscan can help monitor the prevalence and diversity of ARGs in microbial populations and metagenomic samples from animal, human, and environmental settings. The authors intend to constantly update the current version of BacARscan as and when new ARGs are discovered. Executable versions, source codes, sequences used for development and usage instructions are available at (http://www.proteininformatics.org/mkumar/bacarscan/downloads.html) and GitHub repository (https://github.com/mkubiophysics/BacARscan).

Graphical Abstract

Keywords: antibiotic resistance; surveillance tool; monitoring; environmental metagenome; epidemiology; microbial communities
Introduction

Antibiotics—the ‘miracle drugs’—were modern medicine’s revolutionary discovery. However, their frequent and indiscriminate use for clinical, veterinary, agricultural, poultry and other purposes has resulted in bacterial resistance for almost all the significant antibiotic classes [1, 2]. Antibiotic resistance has become a global phenomenon spanning humans, animals and the environment. Hence, regular surveillance of antibiotic resistance genes (ARGs) in microbes and metagenomes from human, animal and environmental sources is vital to understand the epidemiology of ARGs and to foresee the emergence of new antibiotic resistance determinants.

A rapid decline in the cost of high-throughput whole-genome sequencing (WGS) technologies has yielded a plethora of information about ARGs in the microbial population [3–5]. Also, metagenomic studies have added immensely to our knowledge of the microbial ARG pool [6]. As the cost of genome sequencing is decreasing, it is expected that classical microbiological and molecular methods of antibiotic resistance determination might be replaced by WGS-based approaches to microbial ARGs identification methods [7–9]. Antibiotic resistance databases and in silico ARG prediction tools leverage the information generated through classical microbiological and molecular experimental methods to determine antibiotic resistance. This means the efficiency of an in silico tool of microbial ARGs identification methods depends on the quality of data generated by classical microbiological and molecular methods of antibiotic resistance determination. Currently, a significant hindrance in identifying and annotating ARGs from WGS data is the abundance of fragmented genes/genomes in the genome databases (due to incomplete assembly).

In the metagenomic studies of ARGs, the source of genetic material is either the environment, or site of infection or microbiome. The DNA is extracted from the sample and sequenced using the whole-genome shotgun sequencing technique. The fragments are then assembled into contigs. Sometimes a very high genomic diversity may result in a low level of genome coverage that may lead to a small number of contigs and many unassembled sequencing reads (singletons). Thus, an ideal bioinformatics tool, which can detect ARGs in WGS data, should also work effectively on short sequencing reads and fragmented contigs.

Several bioinformatics resources have been developed for cataloguing and characterizing ARGs [10]. The prominent ones are Antibiotic Resistance Genes Database (ARDB) [11, 12], Comprehensive β-lactamase Molecular Annotation Resource (CBMAR) [12], ResFinder [13], Comprehensive Antibiotic Resistance Database (CARD) [14, 15], Resfams [16], Metagenomic Markov models for Antimicrobial Resistance Characterization (Meta-MARC) [17], Antimicrobial Resistance Gene Finder, AMRFinderPlus [18], etc.

ARDB was the first in silico resource that provided a centralized repository to characterize ARGs. ARDB is no longer updated, and its data are incorporated in the CARD. Resfams is a database of hidden Markov models (HMMs) developed using the protein families associated with antibiotic resistance [16]. Meta-MARC is based on hierarchical HMMs, which can predict AMR in metagenomic data (either a short read or a longer assembled contig) into resistance class, group and mechanism. AMRFinderPlus identifies acquired AMR genes and resistance-associated point mutations in protein or assembled nucleotide sequences. CARD identifies and annotates ARGs using BLAST or Resistance Gene Identifier (RGI) modules. A similarity search is performed using BLAST against the CARD reference sequences in the CARD sequence, while the RGI option predicts ARGs based on homology and single nucleotide polymorphism (SNP) models [14, 19]. Several studies have pointed out that BLAST works well in comparing sequences with a high degree of similarity (60% or higher) but does not identify a distant homolog [20, 21]. This limits the capacity of BLAST-based methods to annotate novel ARGs. Because sequence alignment methods like BLAST work well in comparing sequences with a high degree of similarity (60% or higher) but do not identify a distant homologue [20, 21]. Pawlowski et al. [22] reported five ARGs in a cave bacterial isolate, Paenibacillus sp. LC231, which does not have characterized homologues, and CARD could not identify them [22].

One major limitation of ARG databases is their bias towards human pathogens and model organisms, which is understandable because most research efforts are directed towards them [23]. In non-model or non-culturable organisms, remote homologues or highly divergent forms of ARGs of human pathogens and model organisms may be present. This effectively means either the absence or very less number of ARGs of lesser-studied microbes/pathogens. Hence identification of remote homologues or novel ARGs is difficult using the traditional sequence similarity methods [24]. This bias also makes it difficult to identify ARGs in less commonly studied microbes [25].

Also, several resources can identify/characterize resistance only against the prominent/frequently prescribed antibiotics [20]. Thus, we have described a new in silico tool for rapid monitoring, characterization and surveillance of all bacterial ARGs in the present study. Named as Bacterial Antibiotic Resistance scan (BacARScan), this tool has the edge over its predecessors as it can also discern ARGs in short sequencing reads and fragmented contigs. BacARScan can be easily integrated into a user-defined ARG annotation pipeline for the detection of ARG variants in microbial genomes. We also compared BacARScan with other ARG annotation tools: AMRFinderPlus, Meta-MARC, RGI-CARD and Resfams. The results indicated less false-positive (FP) prediction of ARG by BacARScan vis-a-vis other methods.

Material and methods

Overall study design

The complete pipeline of the tool is shown in Fig. 1 with a detailed description.

Dataset collection and cleaning

In the present study, the data on ARGs/proteins were retrieved from various antibiotic resistance databases such as ARG-ANNOT [26], CARD [14], CBMAR [12], INTEGRALL [27], RAC [28], Tetracycline + MLS nomenclature, UCARe [29], Lahey Clinic, Resfams [16], ResFinder [13], HMP [30], LacED [31], MvirDB [32], Institut Pasteur, Patric [33] and published research papers. All ARGs were first grouped based on the mechanisms they employ to inactivate the effect of antibiotics profile. For example, hydrolysis is used to inactivate the β-lactams hence all β-lactamases were grouped; similarly, all antibiotic efflux proteins were grouped. Members of each group were then clustered using BLASTClust with an identity threshold of 90% and coverage of 95%. After removing protein clusters containing less than five sequences or fragmented sequences, we are finally left with 254 protein sequence clusters. Clusters with less than five sequences were removed because a small number of sequences might not provide all mutational information to a profile HMM.
Building protein profile ARGhmm from sequences of antibiotic resistance proteins

All the protein sequences within the 254 clusters were subjected to multiple sequence alignment (MSA) using Muscle 3.8 [34] at default parameters. To check the accuracy of an alignment, we manually checked each alignment and removed sequences that were too divergent [35, 36]. Each MSA was then converted into a profile HMMs using the hmmbuild module of HMMER (http://hmmer.wustl.edu/) (version 3.1) at default parameters [37]. The library of protein profile ARG HMMs was referred to as pARGhmm.

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

In order to develop a tool that can be used directly on genomics/metagenomics sequence data without translating into protein sequences, we created a nucleotide version of the profile HMMs library, named nARGhmm. nARGhmm was created using nucleotide sequences back translated from protein to gene sequences using the Backtranseq (Predicts the potential nucleic acid sequence from which a given peptide sequence originated) program EMBOSS. The procedure followed to build nARGhmm was the same as that followed for pARGhmm. For each antibiotic resistance cluster, we constructed two profile HMMs, one for proteins (pARGhmm) and another for genes (nARGhmm). Collectively, pARGhmm and nARGhmm were named as BacARscan. The schema used for building BacARscan from the set of curated antibiotic resistance proteins (ARPs) and gene sequences is shown in Fig. 1.

Datasets

Dataset I: evaluation dataset

The efficiency of BacARscan was evaluated using a positive and negative dataset. The negative dataset was composed of protein clusters formed during clustering by BLASTclust and had less than five sequences. Since these proteins were not used to create the profile HMMs, they were designated as negative datasets. In order to confirm that the negative dataset is composed of non-ARG sequences, we did gene ontology (GO) term enrichment analysis of negative dataset. The results of enrichment analysis indicated that the most frequently occurring GO terms were not related to AMR, barring a minor fraction showing ‘beta-lactamase activity’ as molecular function and ‘beta-lactam antibiotic catalytic process’ as biological processes (Supplementary Fig. S1). We

Figure 1: The complete workflow depicting the methodology used for the development of profile HMMs-based prediction tool BacARscan.
further confirmed this using RGI-CARD. Out of 65 816 proteins of negative dataset, 6007 (~10%) were characterized as ARG through RGI-CARD out of which 760 were categorized as perfect, 450 as strict and 4797 as loose. It was observed that a majority of ARGs (1574) were characterized as β-lactamases (perfect: 186, strict: 94 and loose: 1294). Details of the classification are shown in Supplementary Table S1. We feel this might have happened because β-lactamase is a highly diverse group of ARG, and the negative dataset has those β-lactamases that are less abundant in nature. Results of RGI-CARD and GO-term enrichment analysis indicated that the negative dataset is mostly (90%) composed of non-ARG sequences. Sequences used to build the pARGhmm, that is protein clusters that contained five or more sequences, were called positive datasets. The combined dataset of full-length positive dataset sequences and both full-length and partial-negative dataset sequences were used for performance evaluation.

**Dataset II: independent dataset**

For benchmarking, we used two independent datasets consisting of (i) penicillin-binding proteins (PBPs) or DD-peptidase proteins, retrieved from the UniProtKB database on 15 April 22 using keyword search (penicillin-binding proteins). This dataset contains 60 reviewed non-fragmented PBPs whose existence was established at the protein level. (ii) Non-antibiotic resistant bacterial efflux (non-ARE) proteins were collected from our earlier work [38]. The final benchmark independent datasets comprised 60 PBPs and 389 non-ARE protein sequences.

**Dataset III: annotation of ARGs in different strains of ESKAPE pathogens**

To assess the ability of BacARscan to discern the diversity of ARGs among different ESKAPE pathogens (Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa and Enterobacter species), we annotated the ARGs present in five strains of an ESKAPE pathogen using BacARscan and compared its performance with other in silico resources, namely RGI-CARD and Resfams. For annotation, we used proteomes of those ESKAPE pathogens whose genomes were completely sequenced, with complete chromosomal and plasmid protein sequences available at National Center for Biotechnology Information (NCBI). Detailed information about ESKAPE pathogens used in this study is mentioned in Supplementary Table S2. During evaluation, the best-hit approach (lowest e-value and highest HMM score) was used for HMM-based methods (BacARscan and Resfams). But in the case of RGI-CARD, only perfect (e-value = 0, sequence identity ~99–100% and bit score above 500) and strict hits (e-value =~1e-80–1e-90, sequence identity ~90–65% and bit score at least 500) were considered.

**Dataset IV: annotation of ARGs using short sequence reads**

The performance of BacARscan was also evaluated on a dataset of short reads generated from the back-translation of protein sequences of positive (used for building profile HMMs: ≥5 cluster sequences) and negative (not used for building profile HMMs: <5 cluster sequences) evaluation dataset. The length of each read was 100 nucleotides (nt), and from each gene sequence, 20 short sequences were generated for evaluation.

**Dataset V: validation dataset**

Simulated short-read data of ARG.

This dataset was used to benchmark BacARscan vis-a-vis other ARG prediction and annotation methods. It was generated using the gene sequence of the ‘protein homolog’ model of the CARD (date: 29 July 2022). The ‘protein homolog’ model contains 4422 ARG sequences that do not include mutation as a determinant of resistance. Using InSilicoSeq [https://in silhouette.readthedocs.io/][39], short reads of 151-nt length at 20× coverage were simulated, imitating the NovaSeq platform of Illumina. A total of 100 000 reads were randomly picked for benchmarking.

**Simulated non-ARG short-read data.**

We also created a non-ARG short-read dataset using the complete genome of a probiotic strain of Enterococcus faecium Strain T-110 (NCBI Genome Accession Number: CP006030). As per Natraj and Parani, this strain, called T-110, does not contain ARGs against all major clinically relevant antibiotics [40]. We used the genome of T-110 to simulate a database of 2 million short reads of 151 bases using InSilicoSeq [https://in silhouette.readthedocs.io/], imitating the NovaSeq platform of Illumina. The comparative evaluation was carried out among BacARscan, Meta-MARC and ResFinder using randomly selected 2M reads.

**Performance measure parameters**

During an evaluation, only predictions with e-values ≤1e-6 were considered significant because maximum precision, recall and F-score were observed at this value (details in comparative evaluation of prediction efficiency of pARGhmm and nARGhmm). To evaluate the efficiencies of both nARGhmm and pARGhmm, we classified the prediction results into four categories: true positives and true negatives (TP and TN, respectively), FP and false negatives FN, respectively. TP and TN represented the correct predictions, while FP and FN represented incorrect predictions. When an ARG sequence was rightfully associated/predicted to its actual class, the prediction was classified as TP, and when a non-ARG was predicted as non-ARG, it was classified as TN. When an ARG and a non-ARG were predicted as non-ARG and ARG, the prediction results were labelled FN and FP, respectively. Using this classification criterion, we calculated the precision, recall, F-score, false-positive rate (FPR) and true-negative rate (TNR) or specificity 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 FP prediction. Precision was calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100$$  \hspace{1cm} (1)

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 very low false predictions. It was calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \times 100$$ \hspace{1cm} (2)

iii) F-measure or F1 or F-score: In order to balance the precision and recall values due to the unequal composition of positive and negative datasets, F-score was also calculated [41]. It is the harmonic mean of precision and recall, and F-score was calculated as:
Results and discussion
Performance assessment of BacARscan (pARGhmm and nARGhmm)

Evaluation of protein dataset
To assess the performance of pARGhmm, we used an evaluation dataset (positive and negative). For this, two different approaches were used to evaluate the performance, that is the top hit and multiple top hits (such as top 3rd, top 5th, top 7th hits, etc.) at different e-value and prediction score cut-offs. When the top hit (based on an e-value and prediction score) was used to evaluate the prediction capability, 228 TP and 26 FP predictions were found (Table 1). Furthermore, the performance of pARGhmm was also evaluated using multiple search results based on the lowest e-value and highest prediction score, and the consensus approach decided the outcome. With the majority approach, the maximum performance was achieved when the top five hits were used to evaluate the performance. The best precision and F-measure were 92.12% and 95.90%, respectively. Overall, the assessment results showed that the prediction performance increased till the decision was made based on the top five hits (based on e-value and prediction score); afterwards, the performance started decreasing. The consistency in search efficiency of each pARGhmm was evaluated using the leave-one-out cross-validation (LOOCV) approach. During LOOCV, one sequence was excluded from the process of model building. Then the performance of the model was evaluated against a dataset containing (i) excluded sequences, (ii) sequences used for building 253 HMMs and (iii) ARP sequences not used for building of pARGhmm. Using this approach, a perfect consistency (~100%) was observed in the performance of the majority of profile HMMs (Supplementary Fig. S2).

Evaluation of short sequencing reads
The performance of nARGhmm was also evaluated on a dataset consisting of short sequences of 100-nt length. The short sequence library was generated from used and non-used sequences for building the profile HMMs, which constitute the positive and negative evaluation datasets. Using the first hit for evaluation, the precision and F-measure values were 90.94 and 95.25, respectively (Table 1). When the performance of nARGhmm was evaluated using the majority approach, the best performance was achieved when the top five hits were used to quantify the performance.

Thus the comparative evaluation of prediction efficiency of nucleotide and protein modules of BacARscan (nARGhmm and pARGhmm) indicated a comparable performance in terms of precision and F-measure (Table 1). The results also indicated the comparable performance of BacARscan on fragmented and full-length datasets, indicating that BacARscan is an efficient tool for finding ARGs in both protein and gene datasets.

Assessment of BacARscan ARG prediction capability vis-a-vis other methods
We evaluated BacARscan with other existing methods using three datasets which include (i) full-length gene/protein homologous non-ARG sequences, (ii) nucleotide short reads of ARGs and (iii) nucleotide short reads of non-ARGs.

Benchmarking of BacARscan on homologous non-ARPs/ARGs
To verify the capability of BacARscan to correctly differentiate ARPs/ARGs from homologous non-ARGS/proteins, we constructed an independent dataset consisting of PBPs and non-antibiotic efflux proteins [38] (non-ARE). Both PBPs and β-lactamases belong to the superfamily of serine penicillin-recognition enzymes and have similar conserved protein folds [42, 43]. It is pertinent to mention that PBPs and non-ARE proteins were not used to construct the HMMs. As done by the previous researchers [13–16], during benchmarking, we also

### Table 1: Performance of BacARscan (pARGhmm and nARGhmm)

| Parameters | Modules       | TP    | FP    | Precision (%) | F-measure (%) | TP    | FP    | Precision (%) | F-measure (%) |
|------------|---------------|-------|-------|---------------|---------------|-------|-------|---------------|---------------|
| No. of top hits | pARGhmm       |       |       |               |               |       |       |               |               |
| 1          |               | 228   | 26    | 89.76         | 94.60         |       |       |               |               |
| 3          |               | 229   | 25    | 90.15         | 94.82         |       |       |               |               |
| 5          |               | 234   | 20    | 99.12         | 95.90         |       |       |               |               |
| 7          |               | 233   | 21    | 91.73         | 95.68         |       |       |               |               |
| 9          |               | 232   | 22    | 91.33         | 95.47         |       |       |               |               |
| 11         |               | 209   | 45    | 82.28         | 90.28         |       |       |               |               |
| 13         |               | 182   | 72    | 71.65         | 83.48         |       |       |               |               |
| 15         |               | 158   | 96    | 62.20         | 76.69         |       |       |               |               |
|            | nARGhmm       |       |       |               |               |       |       |               |               |
|            |               | 231   | 23    | 90.94         | 95.25         |       |       |               |               |
|            |               | 235   | 19    | 92.51         | 96.11         |       |       |               |               |
|            |               | 237   | 17    | 93.30         | 96.53         |       |       |               |               |
|            |               | 236   | 18    | 92.91         | 96.32         |       |       |               |               |
|            |               | 240   | 14    | 94.48         | 97.16         |       |       |               |               |
|            |               | 241   | 13    | 94.88         | 97.37         |       |       |               |               |
|            |               | 240   | 14    | 94.48         | 97.16         |       |       |               |               |
|            |               | 238   | 16    | 93.70         | 96.74         |       |       |               |               |

The overall assessment of BacARscan using both versions of profile HMMs: Scanning of ARPs and ARGs using both versions of profile HMMs (pARGhmm and nARGhmm) against evaluation dataset (positive and negative protein sequences dataset) and short-read sequences independent dataset. The first column shows a couple of alternative top hits like top 1st, 3rd, 5th, 7th hits, etc. TP and FP are the calculation of accuracy of our prediction method. Precision known as exactness was calculated with the above predicted values of TP and FP rate, respectively. Recall and specificity were calculated and we found 100 and 0 in all the cases. F-measure known as harmonic mean was calculated with the help of all considered performance measures.
adopted the best-hit approach with an e-value cut-off of 1e-20. During analysis, out of a total of 60 PBP homologous gene/protein sequences of independent dataset, BacARscan, AMRFinderPlus, Meta-MARC, RGI-CARD, Resfams predicted 6, 12, 9, 15 and 4 as β-lactamases ARGs. In contrast, in non-ARE gene/protein sequences of independent dataset, BacARscan, AMRFinderPlus, Meta-MARC, RGI-CARD, Resfams predicted 23, 37, 26, 91 and 24, out of the 389 non-ARE genes/proteins as efflux-related ARGs (Table 2). In terms of FPR and TNR, though Resfams outperformed all other methods when using FBP, the performance of BacARscan was also comparable. On the other hand, on the non-ARE dataset, the performance of BacARscan was maximum, with 5.9% FPR and 94.08% TNR (Table 2). Overall, results revealed that BacARscan could fairly discriminate between the ARGs and non-ARG consisting of homologous proteins.

### Comparative prediction performance of BacARscan on ARG short-read dataset

Out of 100 000 short reads, BacARscan, Meta-MARC and ResFinder predicted 58 703, 69 294 and 88 831, respectively, as belonging to ARG sequences at an e-value threshold of 1e-6 (Table 3).

**BacARscan has a lower FPR on non-ARGS short-reads dataset than other algorithms**

At the default e-value threshold of 1e-6, BacARscan, Meta-MARC and ResFinder predicted 3979, 22 331 and 1912 short reads, respectively, to belong to ARGs (Table 4). These reads belong to 19, 56 and 5 ARG genes of T-110, respectively, and 3, 18 and 3 ARGs, respectively, in T-110. When a more stringent e-value cut-off of 1e-20 was used, as done in earlier work [13, 14, 16], the number of reads predicted by BacARscan, Meta-MARC and ResFinder was reduced to 238, 9034 and 1648, respectively. When a highly stringent e-value of 1e-50 was used, the no reads were predicted by BacARscan and Meta-MARC as reads of ARG genes, but ResFinder still predicted 1500 reads as ARG. At all e-values, the FPR was <1% except ResFinder at an e-value of 1e-6. This means the rate of TN prediction was nearly 99%.

### Annotation of ARGs in ESKAPE pathogens

The ESKAPE pathogens are globally recognized as a leading cause of nosocomial multidrug-resistant bacterial infections.

The results of ARG annotations by BacARscan revealed a predominance of antibiotic efflux-related ARGs, followed by ARGs encoding β-lactamases or acetyltransferases in five strains of the ESKAPE pathogens (Fig. 2). Similar results were also reported in an earlier published study [44].

### Description of the BacARscan web server

The BacARscan tool is available as a web server at [http://proteinformatics.org/mkumar/bacarscan], where the users can upload the gene/protein sequences for prediction and annotation. The output of BacARscan contains the complete annotation of the query sequence. BacARscan can process a maximum of 10 sequences in one go. For whole-genome/metagenomic/proteome scale annotation, a standalone version is required or available in the web server’s download section or GitHub repository (https://github.com/mkubiophysics/BacARscan).

### Table 3: Performance of BacARscan and other off-the-shelf tools in predicting antibiotic resistance, an external test set of ARG short-read data

| E-value Threshold | No. of simulated reads | No. of reads predicted (hits found) |
|-------------------|------------------------|-----------------------------------|
| BacARscan         | Meta-MARC              | ResFinder                         |
| 1e-6              | 100 000 AR             | 58 703                            | 69 294                           | 88 831 |
| 1e-3              | short reads            | 66 802                            | 77 667                           | 89 580 |
| Default (10)      |                        | 78 680                            | 89 778                           | 99 875 |

### Table 4: Performance of BacARscan and other off-the-shelf tools in predicting antibiotic resistance in an external test set of non-ARG short-read data

| E-value Threshold | Tools              | No. of reads predicted (hits found) | Unique ARGs | FPR (%) | TNR (%) |
|-------------------|-------------------|-----------------------------------|-------------|---------|---------|
| 1e-6              | BacARscan         | 3979                              | 19          | 0.20    | 99.80   |
|                   | Meta-MARC         | 22 331                            | 56          | 1.12    | 98.88   |
|                   | ResFinder         | 1912                              | 5           | 0.10    | 99.90   |
| 1e-20             | BacARscan         | 238                               | 3           | 0.02    | 99.98   |
|                   | Meta-MARC         | 9034                              | 18          | 0.46    | 99.54   |
|                   | ResFinder         | 1648                              | 3           | 0.09    | 99.91   |
| 1e-50             | BacARscan         | 0                                 | 0           | 0       | 0       |
|                   | Meta-MARC         | 0                                 | 0           | 0       | 0       |
|                   | ResFinder         | 1500                              | 3           | 0.08    | 99.92   |
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 quick and cost-effective method of antibiotic resistance determination. BacARscan is an effort to provide an additional way to identify the ARGs in -omics (proteomics/genomics and metagenomic) datasets. BacARscan can also be combined with 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 the future, we will extend it to the new ARGs or families. We hope that BacARscan will help predict ARGs and help in the progress of studies related to antimicrobial resistance (AMR).

Concluding remarks

In the present study, we have described a novel ARG annotation in silico resource named BacARscan, which can be used for rapid monitoring, surveillance and characterization of antibiotic resistance determinants in both genomics and proteomic datasets. Comparison with other in silico resources like AMRFinderPlus, Meta-MARC, Resfams and CARD revealed that BacARscan’s ability to discern ARGs in -omics datasets was much more significant than its predecessors. One of the most notable improvements of BacARscan over other ARG annotation methods is its ability to work on genomes and short-reads sequence libraries with equal efficiency and without any requirement for assembly of short reads. Thus, BacARscan would help monitor the prevalence and diversity of ARGs in microbial populations and metagenomic samples from animal, human and environmental settings. Though the performance of BacARscan on different datasets looks promising, we understand that without a constant update, it would not remain helpful to the scientific community. The authors intend to constantly update the current version of BacARscan as and when new ARGs are discovered.

Table 5: Comparative ARG predictions analysis of developed method BacARscan with existing methods Resfams and RGI-CARD in various strains of ESKAPE pathogens

| Organism name | Strain designation | Total protein count | Number of predicted ARGs |
|---------------|--------------------|---------------------|--------------------------|
| Acinetobacter baumannii | 6200 LAC-4 ORAB01 D36 AR_0056 | 3708 3638 3764 3786 | 81 88 83 86 |
| Klebsiella pneumoniae | Kpn555 KP69 KPN1H1 MS6671 | 5429 5661 5575 5714 | 89 99 100 97 |
| Staphylococcus aureus | Mu50 N315 ST398 TCH1516 USA300_FPR3757 | 2844 2739 2732 2900 2916 | 64 65 65 66 63 |
| Pseudomonas aeruginosa | EB130952 CR1 K34-7 PA7790 | 6496 5533 6480 6580 | 99 97 96 80 |
| Enterococcus faecium | DO E0139 E4227 E6043 T110 | 3114 2383 2345 2845 2539 | 70 62 64 63 64 |
| Enterobacter spp. | 638 FY-07 CRENT-193 ECN17 | 4358 4741 4909 5082 | 90 86 107 98 |
| Total | | | | 128 305 2537 2182 775 |

The number of ARGs predicted by BacARscan, Resfams and RGI-CARD The e-value threshold: 1e-6 was used for BacARscan and Resfams, while in the case of RGI-CARD, perfect and strict cut-offs were used for the selection of the hits, respectively.
Supplementary data

Supplementary data is available at Biology Methods and Protocols online.

Data and software availability

The tools and all datasets used for evaluation are freely accessible without any restriction on the download page of the web server: http://proteininformatics.org/mkumar/bacarscan/downloads.html

Acknowledgements

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

Authors’ contributions

D.P. collected and organized the data and developed the web interface. D.P., B.K. and M.K. analysed the results. M.K. conceived the idea and did overall supervision of the work. D.P., N.S. and M.K. wrote and reviewed the manuscript.

Funding

The work was carried out using the resources funded by Indian Council of Medical Research projects [ISRM/12(33)/2019 and VIR (25)/2019/ECD-1]. D.P. is supported by the Department of Science and Technology, Government of India under the INSPIRE Program [grant number: DST/INSPIRE 03/2015/003022]. N.S. was supported by the Council of Scientific Research under the Pool Scientist Scheme [grant number: 13(9089-A)/2019- POOL]. B.K. supported by the Council of Scientific Research under the Pool Program [grant number: DST/INSPIRE 03/2015/003022]. M.K. was supported by the Indian Council of Medical Research projects [ISRM/12(33)/2019 and VIR 26. Gupta SK, Padmanabhan BR, Diene SM et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 2014; 58:212–20.

Conflict of interest statement. There are no conflict of interest for all authors.

References

1. Baquero F, Tedim AP, Coque TM. Antibiotic resistance shaping multi-level population biology of bacteria. Front Microbiol 2013; 4:15.
2. Chandra H, Bishnoi P, Yadav A et al. Antimicrobial resistance and the alternative resources with special emphasis on plant-based antimicrobials-A review. Plants 2017; 6:16.
3. Azarian T, Cook RL, Johnson JA 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 2015; 36:777–85.
4. Dominguez SR, Anderson LJ, Kotter CV et al. Comparison of whole-genome sequencing and molecular-epidemiological techniques for Clostridium difficile strain typing. J Pediatr Infect Dis Soc 2016; 5:329–32.
5. Kinnevey PM, Shore AC, Mac Aogáin M 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. J Clin Microbiol 2016; 54:445–8.
6. Schmieder R, Edwards R. Insights into antibiotic resistance through metagenomic approaches. Future Microbiol 2012; 7:73–89.
7. Orincuc EA, Likotrafiti E, Alvarez-Molina A et al. The present and future of whole genome sequencing (WGS) and whole metagenome sequencing (WMS) for surveillance of antimicrobial resistant microorganisms and antimicrobial resistance genes across the food chain. Genes 2018:9:268.
8. Waddington C, Carey ME, Boinett CJ et al. Exploiting genomics to mitigate the public health impact of antimicrobial resistance. Genome Med 2022; 14:15.
9. Köser CU, Ellington MJ, Peacock SJ. Whole-genome sequencing to control antimicrobial resistance. Trends Genet 2014; 30:401–7.
10. Xavier BB, Das AJ, Cochrane G et al. Consolidating and exploring antibiotic resistance gene data resources. J Clin Microbiol 2016; 54:851–9.
11. Liu B, Pop M. ARDB–antibiotic resistance genes database. Nucleic Acids Res 2009; 37:D443–47.
12. Srivastava A, Singhal N, Goel M et al. CBMAR: a comprehensive β-lactamase molecular annotation resource. Database (Oxford) 2014; 2014: bau111.
13. Zankari E, Hasman H, Cosentino S et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 2012; 67:2640–4.
14. McArthur AG, Waglechner N, Nizam F et al. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 2013; 57:3348–57.
15. Alcock BP, Raphenya AR, Lau TTY et al. CARD 2020: antibiotic Resistome Surveillance with the comprehensive Antibiotic Resistance Database. Nucleic Acids Res 2020; 48: D517–25.
16. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J 2015; 9:207–16.
17. Lakin SM, Kuhnle A, Alpanahi B et al. Hierarchical hidden Markov models enable accurate and diverse detection of antimicrobial resistance sequences. Commun Biol 2019; 2:294.
18. Feldgarden M, Bruver V, Gonzalez-Escalona N et al. AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 2021; 11:12728.
19. Jia B, Raphenya AR, Alcock B et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res 2017; 45: D566–73.
20. Arango-Argoty G, Garner E, Pruden A et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome 2018; 6:23.
21. Kleinheinz KA, Joensen KG, Larsen MV. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and virulence genes in bacteriophage and prophage nucleotide sequences. Bacteriophage 2014; 4: e27943.
22. Pawlowski AC, Wang W, Koteva K et al. A diverse intrinsic antibiotic resistome from a cave bacterium. Nat Commun 2016; 7: 13803.
23. Boolchandani M, D’Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet 2019; 20: 356–70.
24. McArthur AG, Tsang KK. Antimicrobial resistance surveillance in the genomic age. Ann N Y Acad Sci 2017; 1388: 78–91.
25. Yelin I, Kishony R. Antibiotic resistance. Cell 2018; 172: 1136–1136 e1.
26. Gupta SK, Padmanabhan BR, Diene SM et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 2014; 58: 212–20.
27. Moura A, Soares M, Pereira C et al. INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. Bioinformatics 2009;25:1096–8.

28. Tsafnat G, Copty J, Partridge SR. RAC: repository of antibiotic resistance cassettes. Database (Oxford) 2011;2011:bar054.

29. Saha SB, Uttam V, Verma V. “U-CARE: user-friendly comprehensive antibiotic resistance repository of Escherichia coli. J Clin Pathol 2015;68:648–51.

30. Human Microbiome Project Consortium. A framework for human microbiome research. Nature 2012;486:215–21.

31. Thai QK, Bös F, Pleiss J. The Lactamase Engineering Database: a critical survey of TEM sequences in Public Databases. BMC Genomics 2009;10:390.

32. Zhou CE, Smith J, Lam M et al. MvirDB–a Microbial Database of protein toxins, virulence factors and antibiotic resistance genes for Bio-defence applications. Nucleic Acids Res 2007;35:D391–4.

33. Wattam AR, Abraham D, Dalay O et al. PATRIC, the Bacterial Bioinformatics Database and analysis resource. Nucleic Acids Res 2014;42:D581–91.

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

35. Chiu JKH, Ong RT-H. ARGDIT: a Validation and Integration Toolkit for Antimicrobial Resistance Gene Databases. Bioinformatics 2019;35:2466–74.

36. Papp M, Solymosi N. Review and comparison of Antimicrobial Resistance Gene Databases. Antibiotics (Basel, Switzerland) 2022;11:339.

37. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 2011;39:W29–37.

38. 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 2020;10:9287.

39. Gourlé H, Karlsson-Lindsjö O, Hayer J et al. Simulating illumina metagenomic data with InSilicoSeq. Bioinformatics 2019;35:521–2.

40. Natarajan P, Parani M. First complete genome sequence of a probiotic Enterococcus Faecium Strain T-110 and its comparative genome analysis with pathogenic and non-pathogenic Enterococcus Faecium genomes. J Genet Genomics 2015;42:43–6.

41. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol 2020;2(1):37–63. http://arxiv.org/abs/2010.16061.

42. Knox JR, Moews PC, Frere J-M. Molecular evolution of bacterial β-lactam resistance. Chem Biol 1996;3:937–47.

43. Meroueh SO, Minasov G, Lee W et al. Structural aspects for evolution of beta-lactamases from penicillin-binding proteins. J Am Chem Soc 2003;125:9612–8.

44. Brooks LE, Ul-Hasan S, Chan BK et al. Quantifying the evolutionary conservation of genes encoding multidrug efflux pumps in the ESKAPE pathogens to identify antimicrobial drug targets. mSystems 2018;3:1–9.