Rocker Models for Reliable Detection and Typing of Short Read Sequences Carrying β-lactamases

Si-Yu Zhang
Georgia Tech: Georgia Institute of Technology https://orcid.org/0000-0001-6701-5790
Brittany Suttner
Georgia Tech: Georgia Institute of Technology
Luis Rodriguez-R
Georgia Tech: Georgia Institute of Technology
Luis Orellana
Max-Planck-Institut fur Marine Mikrobiologie
Jessica Rowell
Centers for Disease Control and Prevention
Hattie Webb
Centers for Disease Control and Prevention
Amanda Williams-Newkirk
Centers for Disease Control and Prevention
Andrew Huang
Centers for Disease Control and Prevention
Konstantinos Konstantinidis (kostas.konstantinidis@gatech.edu)
Georgia Institute of Technology https://orcid.org/0000-0002-0954-4755

Research

Keywords: β-lactamases, Short reads, ROCker models, High F1 score

Posted Date: November 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-113339/v1

Abstract

Background: Resistance genes encoding β-lactamases (BLs) confer resistance to the widely prescribed antibiotic class, β-lactams. Therefore, the prevalence of BL genes in clinical or environmental samples is important for assessing the public health risk and the spreading of these genes into pathogens. However, identification of genes encoding BLs from short read metagenomes remains challenging due to the high frequency of shared amino acid functional domains and motifs in proteins encoded by BL genes and related, non-BL gene sequences. Accordingly, divergent BL homologs can be frequently missed during similarity searches, which has important practical consequences for monitoring antibiotic resistance.

Results: To address this limitation, we built ROCker models that targeted either broad classes (e.g., class A, B, C and D) or individual families (e.g., TEM) of BLs and challenged them with mock 150 bp and 250 bp read data sets of known composition. ROCker identifies most-discriminant bit score thresholds in sliding windows along the sequence of the target protein sequence and hence, can account for non-discriminative domains shared by unrelated proteins. BL ROCker models showed a 0% false positive rate (FPR), a 0% to 4% false negative rate (FNR), and a up to 50-fold higher F1 score [2*precision*recall/(precision+recall)] compared to alternative methods, such as similarity searches using BLASTx with various e-value thresholds and BL Hidden Markov Models, or specialized tools for this purpose like DeepARG, ShortBRED and AMRFinder. The ROCker models and the underlying protein sequence reference data sets and phylogenetic trees for read placement are freely available through http://enve-omics.ce.gatech.edu/data/rocker-bla.

Conclusions: Our results showcased the reliable detection and typing of short read sequences carrying BLs by ROCker models. Application of these BL ROCker models on metagenomics, metatranscriptomics as well as high-throughput PCR gene amplicon data should facilitate the reliable detection and quantification of BL variants encoded by environmental or clinical isolates and microbiomes, and more accurate assessment of the associated public health risk compared to the current practice.

Background

Genes encoding β-lactamases (BLs) confer resistance to a widely prescribed class of antibiotics, the β-lactams [1]. These β-lactam-degrading enzymes represent a major public health threat due to their ability to inactivate clinically important β-lactams and rapid dissemination of β-lactamase genes among pathogenic microorganisms [2, 3]. Mobilization of BL genes from environmental organisms into pathogens additionally occurs [4], since β-lactams are naturally produced compounds [5–7], and is also relevant for public health. Currently, more than 1800 BL variants have been described [8]. BL variants are classified into four classes based on conserved active-site amino acid motifs (Ambler classification) [9] or multiple classes based on functional characteristics and substrate/inhibitor profiles (Bush-Jacoby-Medeiros classification) [10]. According to the Ambler classification system, BL variants are classified into four molecular classes: A, B, C and D [9]. Among the Ambler classes, classes A, C and D are serine BLs, which hydrolyze β-lactams by forming an acyl enzyme intermediate through an active site serine,

whereas class B enzymes are metallo-β-lactamases (MBLs) and utilize at least one active-site zinc ion to hydrolyze their substrates [11].

Assessing presence of β-lactamases and their specific variants with next generation sequencing (NGS) technologies in isolate genomes and metagenomes could help guide treatment selection and enable improved surveillance of BL in both clinical and environmental settings (12, 13, 14). However, identifying short reads carrying BL genes remains challenging due to the frequent sharing of amino acid functional domains and motifs between proteins encoded by target antimicrobial resistance genes (ARGs) like BL genes and non-BL or other (non-ARG) gene sequences [2, 3]. Full-length gene sequences assembled as part of genomes or metagenomes generally represent less of a problem with this respect, although the high sequence divergence often observed between and within BL classes represents another major challenge for identification of short read or full-length gene sequences carrying BLs [8, 12]. For instance, distinct BLs, even of the same class, frequently share less than 40% amino acid identity across their sequences, which is problematic for detecting homologous sequences [13].

Identification of ARG-carrying reads typically relies on sequence similarity searches against curated databases such as ResFinder [14], The Comprehensive Antibiotic Resistance Database (CARD) [15], Antibiotic Resistance Gene Database (ARDB) [16] and UniProt (Universal Protein Resource) databases, or the use of ARGs-OAP and DeepARG tools [17–19]. Due to the continuous release of genomes and metagenomes carrying novel BL variants [5], these curated databases are not updated in a timely fashion. Further, BL sequences are frequently annotated inconsistently or erroneously in the public domains [20], meaning that sequences remain unassigned to a family or class, or sequences that are related but not BLs such as lactam-binding (but not hydrolyzing) sequences are identified as BLs. For these reasons, as well as the technical challenges mentioned above related to shared amino acid domains and motifs, the tools and databases available are far from ideal in detecting BL encoding genes from metagenomic samples, especially for those sequences carried by short reads [17–19]. Hence, a reliable approach to detect BLs including novel and divergent variants in short read (unassembled) metagenomic data sets is needed. It is important to note that these issues are also relevant for amplicon sequencing data sets such as those that result from PCR assays with broad-specificity BL primers. In particular, it remains challenging to distinguish between BL sequences (specific priming) and non-BL sequences produced in such amplicon data sets due to non-specific priming [21, 22]. Regions of high sequence identity also often make it difficult to distinguish accurately between multiple BLs in both amplicon and shotgun short read metagenome data [23, 24].

To address the technical limitations in detecting BLs in short-read metagenomic shotgun or amplicon sequences, we built ROCker models for BLs and evaluated them with mock 150 bp and 250 bp read data sets of known composition. Instead of relying on fixed e-value thresholds for the whole alignment, as is typically the case in similarity searches for detecting ARG-carrying sequences which can return an unknown number of false positives and negatives, ROCker identifies position-specific, most-discriminant bit score thresholds in sliding windows along the sequence of the target protein sequence (e.g., BL) using the receiver operating characteristic (ROC) curve. Hence, ROCker can account for non-discriminative

domains shared by unrelated proteins [25]. We have previously reported that ROCker often shows more than a 50-fold lower false discovery rate (FDR) when compared to the common practice of using fixed e-values or Hidden Markov Model (HMM)-based searches for genes of the microbial nitrogen cycle [25]. Here we show that our ROCker models can reliably detect and type BLs in short-read metagenomes with similar difference in FDR and a high accuracy (F1 score) compared to alternative approaches (F1 = ~ 1 vs 0.02–0.86). Specifically, we built ROCker models for all BL classes, i.e., classes A, B, C and D of the Ambler classification system as an example of the between-class resolution that ROCker can provide. We also built ROCker models for TEM (Temoniera β-lactamases, in class A), as an example of the within-class resolution of a highly conserved family of BLs that ROCker models can provide. To avoid repetition of our methodology and results, we present the results for two representative ROCker models, class D BLs (oxacillinase β-lactamases, OXA) and TEM, in the main text; we have provided the remaining models in the supplementary material.

Methods

Overview of the ROCker workflow to identify BLs in short read data

For a detailed description of ROCker, see [25]. Briefly, during the target model building process, a user-provided list of UniProt IDs of positive (target) and negative (non-target) reference proteins are used to identify the corresponding whole-genome sequences encoding these proteins (for how to select appropriate positive reference sequences, see below). The genome nucleotide sequences are downloaded and Illumina-like short reads are simulated from the sequences using Grinder [26] in order to generate short-read training data sets, while keeping track of which reads carry the target reference genes of interest (target reads; based on UniProt IDs provided) and which reads represent the rest of the genome (non-target reads). The positive full-length references are translated and protein sequences aligned using ClustalΩ [27], and the resulting alignment is used as the database to query both the target and non-target simulated short-read data sets, using BLASTx [28] or DIAMOND [29]. The ROCker pipeline then calculates the most discriminant bit score in short windows across the alignment of the reference sequences that best differentiate target vs non-target reads using ROC curves and produces a plot with summarizing statistics of the accuracy of the model based on the training data set. An example of the plot that also includes the graphical representation of the alignments, the calculated bit score thresholds and the matches obtained is provided in Fig. 1. In general, regions of the protein reference alignment with higher amino acid conservation will have higher bit score thresholds than more divergent regions; high bit scores are also expected when non-target sequence share domains or motifs with target sequences based on the training data sets. The ROCker model is composed of the estimated best bit score thresholds for short reads along the positive protein reference alignment and can subsequently be used to filter the search results of a real (or mock) metagenomic data set against the same reference sequences.

Assess the total diversity of BLs and build a reference phylogeny for validating ROCker results and alternative methods

It is important to note that the most critical step in building a ROCker model, and the only step that is not automated, is the choice of the target (positive) vs non-target (negative) reference sequences [25]. Hence, in the following sections, we describe our work with TEM and class D BLs as examples of the procedures that were followed for developing ROCker models for all classes of BLs to avoid redundancy, focusing on the choice of positive/negative sequences. These models and their results are provided in the Supplementary Materials.

To build a list of TEM and class D BL positive references for input into the ROCker model building workflow, we first verified reference sequences of TEM-like variants and class D BL-like variants collected from ResFinder [14] and the Lactamase Engineering Database (part of the BioCatNet) [30–32] in October 2018. Verified sequences included those that were previously shown to inactivate the corresponding β-lactam based on genetic and/or phenotypic tests or showed conservation of functional domains and high sequence identity to such experimentally determined lactamases based on visual inspection of the alignments. These verified sequences were de-replicated using cd-hit version 4.6.1 [33, 34] at a 90% amino acid identity, which resulted in 1 and 39 representative sequences for TEM-like variants and class D BL-like variants, respectively. These reference sequences were then used to search against larger public databases in order to identify additional homologs and cover the total diversity of the proteins currently available. For instance, the experimentally verified TEM-like variants were searched against the Swiss-Prot and TrEMBL (UniRef90, downloaded October 2018) databases using BLASTp (BLAST + version 2.2.28) [28] with a cut-off of coverage of the reference protein length > 90%. In order to capture more recently deposited sequences and any sequences missed by Swiss-Prot and TrEMBL, we also checked for additional, unique matches in the NCBI's NR database.

Finally, all matched reference sequences were de-replicated at 90% amino acid identity using cd-hit version 4.6.1, and one representative per resulting cluster was aligned with the verified references using MAFFT version 7.407 [35]. See the next section for how positive (target) and negative (non-target) sequences were identified specifically for each target protein family. To describe the phylogenetic diversity of TEM-like variants as well as to map short reads on the phylogeny in order to validate the findings of the different tools evaluated (e.g., detect false positives), the MAFFT alignment was used to build a maximum likelihood tree using RAxML version 7.7.2 [36] with the GTRGAMMA model (Fig. 1). Reads identified by the evaluated tools were mapped to this reference phylogeny as described below. The same approach was used for the class D BL sequences except that the TrEMBL sequences were de-replicated at 50% amino acid identity due to the higher level of diversity among members of this class.

Determine the reference protein sequences for building the ROCker model

Protein sequences of TEM were obtained from the UniProt database and were aligned and manually inspected for the presence of the known functional motifs of TEM, i.e., R^{65}FxxxS^{70}xxK, S^{130}DN and K[TS]G [8]. To more comprehensively cover the diversity of this family, TEM from different taxa were included in the positive reference sequences (Supplementary Table S1) for building the ROCker model. In order to improve the performance of the model, a second list of negative (i.e., non-target) references that represented evolutionarily related antimicrobial resistance proteins with different functions such as SHV (sulfhydryl variable β-lactamase), OXY (oxytoca β-lactamase), CTX-M (cefotaxime β-lactamase) and KPC (*Klebsiella pneumoniae* carbapenemase) (Fig. 2 and Supplementary Table S1) were also included, as suggested previously for increased accuracy when related (non-target) proteins are expected to be present in the query data sets [25]. The ROCker model was built using a DIAMOND search [29] with default settings [25].

More specifically, because the TEM-like variants share high sequence identity (98–99% amino acid identity) to each other (low intra-family diversity), only sequences for a few TEM-like representatives are necessary for the positive reference set. In total, 11 TEM protein sequences from different organisms including *Klebsiella oxytoca*, *Neisseria gonorrhoeae*, *Escherichia coli*, *Salmonella enterica* serotype Typhimurium, *Kluyvera georgiana* and *Proteus mirabilis* were included as positive reference sequences for model building (Supplementary Table S1). Due to the functional domains BLs share with non-target proteins such as lactam-binding (but non-hydrolyzing) proteins, and in order to obtain high, family-level resolution with the ROCker model, we included representatives of non-target proteins as negative reference sequences during model building as suggested previously [25]. Reads from non-target gene sequences carrying the functional domains shared with the target sequences will provide relatively high bit scores, and the use of negative reference sequences lowers the probability of these reads being mistakenly taken for positive (target) reads during the testing step. The negative references (33 sequences in total; Supplementary Table S1) were selected from two sets of sequences: the closely-related SHV family (~ 65% amino acid identity to TEM), and non-annotated protein sequences recovered from UniProt that clustered in-between the TEM and SHV (60–80% amino acid identity to TEM) as shown on the phylogenetic tree (Fig. 2). Protein sequences from the non-TEM families of class A BLs [8], including one OXY, one CTX-M and one KPC, were also included as negative references for more comprehensive coverage of related, non-target sequences. In general, for robust ROCker models, it is important for the negative reference sequence set to evenly cover the phylogenetic diversity of related, non-target proteins (if such proteins exist) in order to capture well the range of bit scores provided by short reads originating from such non-target sequences when searched against the reference positive sequences relative to target reads.

For class D BLs, the alignment of reference sequences was manually verified for the two conserved serine residues (i.e., S^{70} and S^{130}) and the K[TS]G domain [8] as described above for conserved functional motifs of TEM. The non-target references such as DacA (D-alanine transpeptidase), MecR1 (methicillin resistance protein), and BlaR1 (penicillin-binding regulatory protein) were also manually verified to have these conserved serine residues and K[TS]G domain because they are motifs of the penicillin binding

superfamily that includes proteins not conferring antibiotic resistance. As these domains are essential for proper enzyme folding and function but are not specific to antibiotic resistance functions, a phylogenetic approach was used to select the positive and negative references for ROCker model building (see Supplementary Table S2 for files used and Supplementary Figure S1). Sequences that formed a clade with non-target reference sequences were included in the negative set. Sequences that were missing at least one of the three functional residues/domains were removed from model building under the assumption that these may be pseudogenes or homologs that have evolved to have different substrate specificities. The positive sequences for ROCker model building included representatives that had the conserved residues/domains and formed a clade with the proteins verified to be class D BLs (which were also included in model building; Supplementary Table S2).

More specifically, a total of 269 class D BL references were collected from the ResFinder database (downloaded October 2018), resulting in 39 sequences when de-replicated at 90% amino acid identity. These sequences showed high sequence diversity among them (e.g., often showing only ~ 30–40% amino acid identity). One representative sequence from each of the two major clades in the class D BL tree (Supplementary Figure S1) was used as a query sequence against the Swiss-Prot database to find the most similar, experimentally-verified non-BL proteins to use as outgroups and as negative references for the model building (described in Supplementary Table S2). The 39 verified class D BL sequences were also queried against the UniRef90 database to find additional class D BLs and cover the described diversity of this class, which resulted in a total of 1,148 matches with query coverage of the target protein length > 90%. These matches were further de-replicated at 50% identity into 227 protein sequences and the alignments were manually inspected for the conserved domains as described above. A total of 29 sequences lacked at least 1 of the 3 domains and thus, were excluded from model building. Another 44 sequences were excluded because there were no corresponding nucleotide sequences for them in the EMBL database which were needed for nucleotide-based short read placement. Finally, 24 proteins formed clades with the verified outgroups and were used as negative references, while 130 sequences were used as positive sequences for the model building. The list of reference sequences for the class D BL model is provided in Supplementary Table S2.

ROCker models for other BLs

ROCker models for class A, B and C BLs were also developed using the same approach as described above for the class D BLs and are described in detail in the Supplementary Material (Supplementary Figure S2-S8). All models are available on our website (http://enve-omics.ce.gatech.edu/rocker/models#fam:BLA). Users can also upload their short read data sets to the website for online analysis using these ROCker models following the “Search online” link.

Simulated MOCK data sets for testing ROCker models and alternative methods

In order to generate mock data sets of known composition, target BL and non-target gene sequences were collected together with the genome sequences that carried these genes (Supplementary Table S3). The

gene sequences were selected to include high-confidence target or non-target sequences that were not used in the construction of the ROCker or hidden Markov models (HMMs, see below) to better challenge these tools. For each gene, the source genome was recovered through the EBI web services and metagenomes were simulated from all whole genome sequences using Grinder [26] with options: -dc '-~*NnKkMmRrYySsWwBbVvHhDdXx' -md uniform 0.1 -mr 95 5 -rd *ReadLength* uniform 5, where *ReadLength* corresponds to 150 or 250 bp, depending on the tested read length. Each simulation was repeated five times. The resulting reads derived from target or non-target genes were tagged accordingly and concatenated, resulting in two mock data sets: one for 150 bp and one for 250 bp reads. These mock data sets are available online at http://enve-omics.ce.gatech.edu/data/rocker-bla. The mock data sets were used to assess the accuracy of ROCker relative to that of alternative methods for BL detection such as BLASTx searches [28] using fixed e-values (10^{-2}, 10^{-5}, 10^{-10}, 10^{-20} and 10^{-30}) or hmmsearch using HMMer version 3.1 [37] with an e-value < 0.1. The BL HMMs were downloaded from the FunGene [38] and Pfam databases [39], or custom-built (available online at http://enve-omics.ce.gatech.edu/data/rocker-bla) using the same positive reference set that was used to build the ROCker BL model to enable direct comparisons.

Phylogenetic placement of target BL gene-carrying reads

To further validate reads from the mock data sets identified by ROCker and alternative methods, we placed the identified reads by each method onto a phylogenetic reference tree as follows: The amino acid sequences of full-length reference BLs were aligned using MAFFT version 7.407 [35] with default parameters, and were then used as a guide to align their corresponding nucleotide reference sequences with the EMBOSS tranalign tool [40]. The identified BL gene-carrying reads were then added to the nucleotide reference alignments using MAFFT version 7.407 [35] with ‘addfragments’ option and were placed in the corresponding phylogenetic tree (amino acid reference BL sequences) using RAxML with -f v option. The placement of the target BL gene-carrying reads was visualized in iTOL [41] after processing the resulting visualization jplace file with JPlace.to_iToL.rb from the Enveomics Collection [42].

Results

We describe below the results of each step required to build and evaluate ROCker models for TEM-like variants and class D BL-like variants as representative examples of a phylogenetically narrow and diverse protein families, respectively. Illumina-like simulated short-read data sets of known composition (mock) were used in the evaluation.

Assessment of different options for building ROCker models based on training data sets

The best performing ROCker models based on model training data sets constructed from positive and negative (when available) sequences were observed when using DIAMOND (detailed in the “Technical challenge and tips” section below). We used the following definitions to determine the quality of results:

True Positive (**TP**); False Positive (**FP**); False negative (**FN**); Positives (**P**) = TP + FP; Negatives (**N**) = TN + FN; False Negative Rate (**FNR**) = FN/[TP + FN]; False Positive Rate (**FPR**) = FP/P; **Sensitivity** = 1-FNR = TP/[TP + FN]; **Specificity** = 1-FPR = TP/P; and **Accuracy** = [TP + TN]/[P + N]. Sensitivity, Specificity, and Accuracy were 98.8%, 99.2%, and 99.0%, respectively for the 150 bp TEM ROCker model using DIAMOND (Fig. 1A). For the 250 bp TEM ROCker model, the same statistics were 96.5%, 99.3% and 97.9%, respectively (Supplementary Figure S9A). For the ROCker model of class D BLs, the best performance was achieved after removing (trimming) the first 100 residues of the protein alignment that were not specific enough for class D BLs (i.e., the bit score was similar between the target and non-target protein sequences). The sensitivity, specificity and accuracy of the trimmed ROCker models were 99.1%, 99.8% and 99.8%, respectively for the 150 bp model (Fig. 1B) and 97.4%, 99.9% and 99.9%, respectively for the 250 bp model (Supplementary Figure S9B). Hence, these models were used in the subsequent analyses.

Building mock data sets and metrics for evaluating performance

Mock metagenomic-like data sets were constructed to evaluate the built ROCker models based on training data sets and compare results against other tools. In short, the mock data set typically included a couple hundred reads (nucleotide sequences) originating from one or two target BL sequences within a background of medium-to-high genome complexity (see Material and Methods for details), including about a dozen related but non-target sequences; the target BL sequences differed between each model (e.g., TEM vs class D) but the remaining of the data sets were similar. The tools were evaluated for their ability to detect (identify) the target reads and not identify as target the non-target reads (see also below). Specifically, the 150 bp mock data set contained 107 bla_{TEM} and 331 class D BL gene-carrying reads that were in-silico generated from one TEM and two class D BLs reference nucleotide sequences using Grinder [26] whereas the 250 bp mock set contained 98 bla_{TEM} and 316 class D BL gene-carrying reads. The two class D BLs shared 20% amino acid identity with each other, and both have a homolog with ~ 90% amino acid identity but were not themselves present in the reference class D BL tree or in the positive reference sequences used to build the ROCker and hidden Markov models. Only one TEM was included in the mock data set because there was not substantial diversity among TEM protein sequences in UniProt. The non-target related sequences used in the mock data set included one class B BL, two class C BLs that share 47% amino acid identity to each other, a D-alanyl-D-alanine carboxypeptidase (DacA) and a hydroxyacylglutathione hydrolase (GloB) that shared some of the conserved domains with target BLs and showed 10% − 50% sequence amino acid identity to different classes of BLs, in addition to reads derived from the other genes in the genomes that encoded these non-target sequences. In order to increase the complexity of the TEM mock data set, another 10 non-target related references were included that represent different clades of the phylogenetic tree of class A BL variants with one or two representatives each. These non-target TEM sequences shared domains and motifs with TEM families at 30% − 80% sequence amino acid identity and were also used as class A BL target sequences for the class A BL data sets. UniProt identifiers for BL nucleotide sequences and their host genome sequences used to create the

mock data sets are listed in Supplementary Table S3. Mock data sets are also available on our website that hosts the ROCker models.

The performance of ROCker models with 150 bp and 250 bp simulated data sets were evaluated and compared with alternative approaches employing a fixed e-value: either a BLASTx search with e-values < 10^{-2}, 10^{-5}, 10^{-10}, 10^{-20} and 10^{-30} or an hmmsearch using BL HMMs with an e-value < 0.1. Only reads with > 90% of their length carrying the target BL genes were considered as target reads (positive matches); reads carrying a shorter fragment of the target gene sequence were not considered as targets or non-targets. For each method evaluated, each target read was either correctly identified as a target BL of interest (true positive; TP) or incorrectly identified as a non-target sequence (false negative; FN). Non-target reads were those that carried any sequence other than the target BL gene, i.e., a gene sequence for an unrelated or related but functionally distinct protein. For each method evaluated, these reads were classified as false positive (FP) matches if they were incorrectly identified as carrying the target BL; otherwise the reads were considered true negatives (TN). Prediction accuracy of each method was measured as the false negative rate [FNR = FN/(TP + FN))], which identifies the failures to detect target sequences; the false positive rate [FPR = FP/(TP + FP)], which identifies the failures to exclude non-target sequences; precision [TP/(TP + FP)], which represents how reliable the detected reads are; recall [TP/(TP + FN)], which represents how efficient the detection of all target reads is; and the F1 score [2*precision*recall/(precision + recall)], a summary metric that combines precision and recall. Note that the latter metric (F1 score) is more comprehensive, but cannot distinguish different failures in the results, and therefore we report all other metrics. ROCker was more accurate (higher F1 score), in general, in detecting reads carrying target BLs from the simulated mock data sets (150 bp and 250 bp) compared to alternative approaches (Fig. 3 and Supplementary Table S4 for TEM, Fig. 4 and Supplementary Table S5 for class D BLs).

Comparison of ROCker models to alternative approaches

TEM ROCker model

For detection of bla_{TEM} carrying reads, ROCker had perfect precision (1.00), recall (1.00) and F1 score (1.00) for both the 150 bp and 250 bp mock data sets, i.e., no FP and FN reads were called. BLASTx searches had relatively low F1 scores, ranging from 0.02 to 0.86 depending on the e-value threshold applied, and detected many more FP and FN reads (0–9681 and 0–26 at different e-values, respectively) compared to ROCker (Fig. 3). With decreasing e-values from 10^{-2} to 10^{-30} in the BLASTx search (higher stringency), the FPR decreased from 99–0% based for the 150 bp read length data set but a concomitant increase of FNR from 0–24% was also observed. For the 250 bp read length data set, no FN was detected by BLASTx searches, but the FPR was always high, e.g., 99% at 10^{-2} and 61% at the 10^{-30} e-value. The HMMs, either downloaded from the FunGene database or custom-built using the same positive references for ROCker, had similar performance. For instance, the custom-built HHM had an F1 score of 0.22 and FPR of 87% for the 150 bp data set, and 0.13 and 93%, respectively for the 250 bp data. A slightly higher FNR was observed for the TEM HMM downloaded from FunGene compared to the custom-

built HMM, i.e., 1% vs 0% and 4% vs 0% for the 150 bp and 250 bp data sets, respectively. The BL HMM downloaded from the Pfam database is not specific for TEM detection (but targets all BLs) and therefore, had a lower F1 score compared to the TEM-specific HMM, i.e., 0.13 vs 0.22 and 0.08 vs 0.13 for the 150 bp and 250 bp data sets, respectively, with a high FPR (93% vs 87% and 96% vs 93% for the 150 bp and 250 bp data sets, respectively) and FNR (53% vs 0% and 20% vs 0% for the 150 bp and 250 bp data sets, respectively; Supplementary Table S4).

Class D BL ROCker model

For detection of class D BL gene-carrying reads, ROCker had perfect precision (1.00), and nearly perfect recall (0.96 and 0.99 for the 150 bp and 250 bp data sets, respectively) and F1 scores (0.98 and 1.00 for the 150 bp and 250 bp data sets, respectively). In contrast, the F1 scores ranged from 0.02 to 1.00 for BLASTx searches depending on the e-value threshold applied (F1 score = 1.00 at e-value of 10^{-30}), and 0.54 to 0.78 for different HMMs (Fig. 4). A much higher FPR was observed for HMMs and BLASTx searches at an e-value of 10^{-2} compared to ROCker, i.e., 21% – 100% vs 0% for the 150 bp data set, and 33% – 100% vs 0% for the 250 bp data sets. The more stringent e-value thresholds at 10^{-5}, 10^{-10}, 10^{-20} and 10^{-30} for BLASTx searches performed well overall and actually had slightly lower FNR than ROCker, i.e., 0% vs 4% for the 150 bp data set, and 0% vs 1% for the 250 bp data set. The HMM downloaded from the FunGene database performed slightly better than the custom-built HMM, with a higher F1 score; i.e., 0.78 vs 0.72 for 150 bp data set, and 0.75 vs 0.54 for 250 bp data set (Supplementary Table S5). The BL HMM downloaded from Pfam was unable to detect any class D BL gene-carrying reads likely because the model was not trained to include class D BLs despite being described as a general BL HMM.

Phylogenetic placement of identified reads to validate results and assess false positives

bla_{TEM}-carrying reads

In total, 43 amino acid reference sequences including 11 positive TEM references and 32 negative references from other (non-target) clades of class A BLs, such as SHV, LAP, TER, OKP, OXY and CTX-M were used to build a reference phylogenetic tree (Fig. 2). To further validate the results obtained by ROCker and the similarity searches and identify the exact sequence variant carried by the reads, the identified *bla*$_{TEM}$carrying reads by each tool in the mock data sets were placed in the reference phylogeny (Fig. 5). The KPC in the reference set of the ROCker model was excluded from the phylogenetic tree because no corresponding nucleotide sequence that was needed for nucleotide-based short read placement was found in the EMBL database for this protein. The phylogenetic placement of the *bla*$_{TEM}$-carrying reads (n = 107 in total; Fig. 5A) in the 150 bp data set, which were identified as TP by all three approaches (i.e., ROCker, BLASTx search with e-value < 10^{-5}, and a custom-built TEM HMM with e-value < 0.1), showed that a majority of the reads (about 60%) grouped together with the TEM family, as expected. The remaining reads were associated with proteins closely related to TEM (75–78% amino acid identity) that clustered in-between the TEM and SHV family, such as LAP, LEN and OKP or SHV family, or placed on

more ancestral nodes (non-discriminating) of the TEM and SHV lineages (about 20% of total remaining reads), presumably because these reads did not carry enough phylogenetic signal for placement to the tips of the reference tree, e.g., they carried domains shared between target (TEM) and non-target (the rest, non-TEM) protein sequences.

No FN reads were observed for any of three approaches for identifying TEM carrying reads. Further, ROCker detected no FP (non-target) reads while the BLASTx search (e-value < 10^{-5}; Fig. 5B) and the custom-built TEM HMM (e-value < 0.1; Fig. 5C) detected 863 and 743 FP reads, respectively. The phylogenetic placement patterns for these reads were consistent with the reads being FPs, i.e., the majority (89%) of the FP reads were placed in the SHV lineage and on the branches in-between the TEM and SHV (proteins closely related to TEM). A small number of the reads, about 10%, were assigned to the more divergent non-target class A clades, i.e., the OXY and CTX-M. Around 1% of the FP reads mapped to the TEM family due to the fact that the reads sampled a non-discriminatory (highly conserved) part of the alignment (Supplementary Figure S10).

Compared to the 150 bp data set, a lower number of TP reads using all methods (98 target reads) and higher number of FP reads (1,396 by BLASTx and 1,287 by custom-built TEM HMM) were observed for the 250 bp data set (Supplementary Figure S11). The placement patterns for the TP (Supplementary Figure S11A) and FP (Supplementary Figure S11B and S11C) reads in the 250 bp data set were similar to those in the 150 bp data set, except a slightly lower percentage (0.5%) of the FP reads were assigned to the TEM family, as expected for longer reads that are more discriminatory. Overall, the FP reads represented closely related protein sequences to the target sequences but of distinct function that the similarity searches (but not ROCker) were unable to distinguish from the target positive reads.

Class D BL gene-carrying reads

The class D BL gene-carrying reads (150 bp data set) identified by ROCker, BLASTx search (e-value < 10^{-5}) and custom-built class D BL HMM (e-value < 0.1) were placed on the phylogenetic tree shown in Fig. 6A. A combined total of 159 (48%) of the TP reads placed onto the branches of the two verified class D BL references that were most closely related to the protein sequences used in the mock data (i.e. the two largest circles in Fig. 6A). The other 52% of TP reads were scattered around the tree on both target and non-target branches and on more ancestral nodes, presumably because these reads originated from conserved regions that are difficult to distinguish between target and non-target sequences.

Overall, a small number of the class D BL gene-carrying reads was missed in the 150 bp data set by these approaches, i.e., 13 FN reads out of the 331 target reads for ROCker (4% of the total class D BL genes-carrying reads) and 4 FN reads out of the 331 target reads for the custom-built HMM (1% of the total class D BL genes-carrying reads) (Fig. 6B and 6C). The BLASTx search at 10^{-5} e-value did not result in any FN reads. The majority of the FN reads (100% and 75% of the total FN reads for ROCker and custom-built HMM, respectively) were placed on target class D BLs. However, a single FN read in the custom-built HMM results (25% of the FN reads) mapped to a non-target sequence. This read originated from a more conserved region of the gene that was difficult to distinguish between target and non-target sequences.

No FP (non-target) reads were detected by ROCker, and the FP reads detected by BLASTx (e-value < 10^{-5}; 42 reads) and custom-built HMM (255 reads) searches of the 150 bp data set were more dispersed around the tree compared to the TP or FN reads. In both cases, the highest number of FP reads placed on a single non-target gene branch, only had 19% (8 reads) of total FP reads for BLASTx search (Fig. 6D) and 11% (27 reads) of total FP reads for the custom-built HMM (Fig. 6E) placed to distinct non-target genes. Similar placement patterns of the TP, FN and FP reads were obtained for the 250 bp mock data set (Supplementary Figure S12), except a slightly lower number of TP reads (331 vs 316 targets for the 150 bp and 250 bp data sets, respectively) and a higher number of FP reads detected by BLASTx (42 vs 113 FP reads for the 150 bp and 250 bp data sets, respectively) and the custom built HMM were observed (255 vs 540 FP reads for the 150 bp and 250 bp data sets, respectively).

Comparison to other bioinformatics tools

We also evaluated other tools for the detection of target ARGs in shotgun data such as DeepARG [19], ShortBRED [43] and AMRFinder (https://ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/). It should be noted that these tools were designed with slightly different goals in mind. For instance, DeepARG is designed for finding novel, deep-branching (divergent) homologs to known ARG. ShortBRED is optimized for accurate functional profiling of metagenomic samples by focusing on unique motifs or segments of the targeted protein; hence, ShortBRED does not usually capture reads representing shared motifs with other non-targeted protein. Similarly, AMRFinder is tuned for complete sequences, not fragments carrying by individual metagenomic reads. Accordingly, the comparison of ROCker to these other tools was performed mostly to further highlight the distinctive strengths of the ROCker models. Consistent with our expectations, we found that the combined results from the four ROCker models covering all known classes of BLs (class A, B, C and D) had higher precision (over 70 times lower FPR) and recall (over 5 times lower FNR) than DeepARG or AMRFinder. Note also that the trends observed with AMRFinder were similar to those of Blastx reported above, consistent with our expectations since AMRFinder is essentially a Blastx search (with tuned parameters). Likewise, ROCker had substantially better recall and precision than ShortBRED (over 30 times lower FNR and over 50 times lower FPR for both 150 bp and 250 bp data sets) (Supplementary Figure S15). All together, these results further corroborated the advantages of ROCker models on identifying short reads carrying a specific protein of interest (target).

Technical challenges and tips when using a ROCker model

Similarity search tool to use

The first version of the TEM ROCker model was built using the default settings of ROCker; i.e., the reference protein sequences provided were queried against the simulated shotgun data sets using BLASTx [28] with the e-value set to 0.01. Only the best matches for each read of the training data sets were considered further to compile and plot the model using the script BlastTab.best_hit_sorted.pl (Enveomics Collection [42]). Based on the plot of the TEM ROCker model (i.e., validation plot before querying the real metagenome or mock data sets), the relatively low sensitivity (85.3%) of the model was

due to an unexpected drop in the bit score of a highly conserved region of the target sequences (Supplementary Figure S13). After manually inspecting the BLASTp output of the target sequences, it became clear that these reads were assigned a lower bit score than expected due to a BLAST filter for low-complexity sequences. To avoid this problem, either BLASTx with '-seg no' or DIAMOND search can be used instead of the default BLASTx settings. A comparison of the plot statistics revealed that using DIAMOND with settings of 'min score' of 20 and 'sensitive' [25], among several settings evaluated, had the highest sensitivity (98.8%) compared to the default BLASTx (85.3%) or BLASTx with '-seg no' settings (95.6%) and was significantly faster as previously demonstrated [29]. Thus, DIAMOND is currently recommended for searching reads against a reference alignment.

Trimming the alignment

While examining the plot of the class D BL ROCker model, we noticed that the first 100 amino acid residues of the alignment contained a region with low conservation among the target sequences (i.e., > 50% of the alignment columns had gaps and/or low amino acid identity) that was similar to the conservation levels between target and non-target sequences, and thus not useful for diagnostic purposes (i.e., resulted in frequent false positive matches). Trimming the first 100 residues from the alignment for the 150 bp class D BL model improved sensitivity by ~ 0.2% (Supplementary Figure S14A). Trimming the last 440 residues from the MBL models improved sensitivity by 61.5% (Figure S5); thus, removing such non-discriminatory regions of the alignment could improve model performance. The length of such non-discriminatory regions of the protein sequence relative to the read length should also be considered in making a decision about trimming the alignment or not (e.g., if the length of such regions are too short compared to the length of the target sequences or the read length, such as half the read length or shorter, then trimming is not as useful).

Implementing a length correction for reads shorter than expected in ROCker

The alignment statistic used to discriminate true from false positive matches in ROCker is the bit score, which is sensitive to alignment length. Consequently, large variations in read lengths may result in unexpectedly high FNR (for shorter reads) and, occasionally, high FPR (for longer reads when closely related non-target proteins are present). In order to account for such length variations we implemented the option -L in ROCker search and filter actions (v1.5.2+), which corrects the bit scores according to the ratio of expected to observed query read length with an additional penalty; i.e., if a read is half of the expected read length, its observed bit score will be multiplied by a factor of 2 to account for the shorter length minus a penalty to avoid overinflating results from reads shorter than expected. The penalty can be variable, which assumes that the probability of observing a proportionally higher bit score if the reads were of the expected length decays with the difference between expected and observed length until the maximum length correction (-M), termed the "triangle method" in the documentation of ROCker. Alternatively, a fixed fraction of the difference can be used as penalty (-P). In our tests, we observed the best results using the -L option together with a fraction penalty -P 0.5. However, the data sets used in this study don't exhibit large read-length variation and therefore, we did not apply this correction. The

correction is recommended for data sets with substantial variation in read length such as when 20–30% or more of the reads differ in length by more than 25%.

Discussion

While similarity or HMM searches are commonly used to functionally annotate short read metagenomes [28, 44–46], deciding the optimal e-value threshold for the searches remains a challenge [47, 48]. In addition to the well-appreciated factors such as the size of the database and the length of the query sequence used, the sequence diversity of the target sequences is also important in determining the optimal e-value threshold. A particularly good example of the difficulties associated with searches for functional annotation are epitomized by the beta-lactamases which include families of proteins with high sequence diversity, as well as families with little to no sequence diversity. For instance, for the class D BLs, which have a higher level of diversity, similarity searches with a fixed e-value at 10^{-10} showed comparable performance to the ROCker model, with virtually no FP or FN detected in the mock data sets used to test these calling methods (Supplementary Table S5). In fact, ROCker missed a small number (4%) of target reads (FN) that Blastx with e-value at 10^{-10} or 10^{-20} did not miss, presumably because these e-value thresholds happen to be ideal for separating target from non-target reads in this (but not necessarily other) mock data sets. However, for TEM (family level BLs), which are highly similar among themselves, using the same e-value threshold at 10^{-10} for similarity searching resulted in a large number of FP reads (434 and 1055 for the 150 bp and 250 bp data sets, respectively); even a more stringent e-value threshold still resulted in a large number of FP reads, especially for the longer read length data set, i.e., FPR of 61% for the 250 bp data set (Supplementary Table S4). It is also important to note that with a more stringent e-value threshold, the decrease in FPR is accompanied by a proportional increase of FNR for BLAST or HMM-based approaches (Figs. 3 and 4), and the optimal e-value threshold to employ remains typically elusive in most studies due to the lack of mock data sets and/or an *a priori* appreciation of the diversity of the target sequences in the metagenomic data set of interest.

The analysis of mock data sets with BL gene-carrying reads showed that ROCker can effectively sidestep these limitations because it is based on a position-specific calculation of the most discriminant bit score thresholds that do not depend on data set size or target length [25]. Accordingly, ROCker provided much higher F1 scores compared to using similarity or HMM searches with a fixed e-value threshold of 0.1 or below for the same purposes (e.g., Figs. 3 and 4). Thus, ROCker is a particularly powerful approach for proteins such as BLs due to the various diversity levels among sub-families and different sequence lengths [8, 49–51]. It's also important to note that the ROCker-based pipeline described above can be applied to amplicon sequencing data, such as those from high-throughput PCR platforms, to evaluate non-specific primer amplification of non-target sequences and/or type target sequences against a reference phylogeny of the target [52, 53].

Maximum likelihood placement of reads against the reference phylogenetic tree can be a highly accurate method for identification and typing of reads carrying the gene of interest (target), in general. However, this method is not practical for processing a large number of reads (in the thousands or millions) in a

reasonable time. Hence, we employed read placement herein which is computationally tractable as an additional step to the ROCker-based filtering of reads. Phylogenetic placement of the reads further validated the results of ROCker and the high frequency of TP calls, i.e., target reads were mapped to the branches expected based on the reference sequence from which the reads originated (Fig. 5A and 6A). However, in several cases the *bla*$_{TEM}$carrying (Fig. 5A) and non-target-carrying reads (FP reads; Fig. 5B and 5C) were placed on the same branches, rendering it challenging to distinguish between TP and FP reads based on the phylogenetic approach. These reads mostly originated from a highly identical region shared between TEM and closely related (non-TEM) proteins such as proteins from the TEM/SHV clade within class A BLs [8] (Supplementary Figure S10). ROCker was able to identify these regions of the target protein at the modeling compilation step by increasing the bit score threshold of the corresponding windows and thus did not make these FP calls. Thus, ROCker can be advantageous in such cases even over phylogeny-based approaches in terms of not overcalling FP matches in regions of high sequence similarity (low phylogenetic signal) between target and non-target sequences. The approach presented here, i.e., to use ROCker to filter matching reads combined with phylogenetic placement of the identified reads against the tree of the positive and negative reference sequences, takes advantage of the strengths of these two methods in a complementary fashion.

The ROCker models for the four known classes of BLs, i.e., class A (Supplementary Figure S3), B (Supplementary Figure S5 and S6), C (Supplementary Figure S8) and D (Fig. 1B and Supplementary Figure S9B) BLs, and also for BLs from the phylogenetically narrow TEM family (Fig. 1A and Supplementary Figure S9A) are available through http://enve-omics.ce.gatech.edu/rocker/models#fam:BLA for online analysis of short read data sets provided by external users. The TEM and class D BL models described herein represent detailed examples of how to build and test ROCker models for highly related proteins and highly diverse proteins, respectively. Selecting the appropriate positive reference sequences, and when needed, negative references, and manually verifying is the key and most time-consuming step in building robust ROCker models, and we provide examples of how to perform this step and caveats for what to avoid.

Conclusions

Our work demonstrated the reliable detection and typing of short read sequences carrying BLs by ROCker models and the higher accuracy of this approach compared to alternatives for the same purpose and type of data. Therefore, application of these BL ROCker models on metagenomics, metatranscriptomics or high-throughput PCR gene amplicon datasets from various environments should enable the reliable detection and quantification of BL variants encoded by environmental or clinical isolates and microbiomes, while avoiding high frequency false positives calls. Furthermore, the curated ROCker models and reference BL sequences available through our webserver should facilitate the development of new models for additional (phylogenetically narrow) BL families as well as non-BL ARGs. Therefore, the ROCker models presented here substantially expand the toolbox for monitoring antibiotic resistance in clinical or environmental samples and assessing public health risk [54].

Abbreviations

BLs: β-lactamases; MBLs: metallo-β-lactamases; TEM: Temoniera β-lactamases; OXA: oxacillinase β-lactamases; SHV (sulfhydryl variable β-lactamase), OXY: oxytoca β-lactamase; CTX-M: cefotaxime β-lactamase; KPC: Klebsiella pneumoniae carbapenemase; DacA: D-alanine transpeptidase; MecR1: methicillin resistance protein; BlaR1: penicillin-binding regulatory protein; ARGs: antimicrobial resistance genes; NGS: next generation sequencing; CARD: Comprehensive Antibiotic Resistance Database; ARDB: Antibiotic Resistance Gene Database; ROC: receiver operating characteristic; HMM: Hidden Markov Model; TP: True positive; FP: False positive; P: Positives; N: Negatives; FPR: false positive rate; FNR: false negative rate;

Declarations

Acknowledgements

We would thank Janet Hatt for her informative comments and revision for the paper. We also thank the Centers for Disease Control and Prevention and U.S. National Science Foundation for their kind support this project.

Author's contributions

SZ, BJS, LMR and KTK conceived the study. SZ built the class A, class C BLs and TEM ROCker models and evaluated their performance. BJS built the class B and class D BLs ROCker models and evaluated their performance. LMR build the 150 bp and 250 bp mock datasets and the online webserver with the models. SZ, BJS, LMR and KTK wrote the paper. LHO, JLR, HEW, AJW, AH provided data and suggestions on the paper. All authors read and approved the final manuscript.

Funding

This work was supported by funds made available from the Centers for Disease Control and Prevention and in part by the U.S. National Science Foundation under award no. 1759831 (to K.T.K.). B. J. S. was supported by U.S. National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650044.

Availability of data and materials

The ROCker models for class A, B, C and D BLs, and for TEM family are available through http://enve-omics.ce.gatech.edu/rocker/models#fam:BLA. The 150 bp and 250 bp mock data sets are available online at http://enve-omics.ce.gatech.edu/data/rocker-bla.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declaim no conflicts of interest.

Author details

[1]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. [2]School of Ecological and Environmental Sciences, East China Normal University, Shanghai, 200241, China. [3]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA. [4]Department of Microbiology and Digital Science Center (DiSC), University of Innsbruck, Innsbruck, Tyrol, Austria. [5] Max-Planck-Institut für Marine Mikrobiologie, Bremen, Germany. [6]Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, GA, USA. [7]Weems Design Studio, Suwanee, GA, USA.

References

1. Davies J, Davies D. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev. 2010;74(3):417-33.
2. Khan AU, Maryam L, Zarrilli R. Structure, genetics and worldwide spread of New Delhi metallo-β-lactamase (NDM): a threat to public health. BMC microbiology. 2017;17(1):101.
3. Pitout JD, Laupland KB. Extended-spectrum β-lactamase-producing Enterobacteriaceae: an emerging public-health concern. The Lancet infectious diseases. 2008;8(3):159-66.
4. Wolters B, Widyasari-Mehta A, Kreuzig R, Smalla K. Contaminations of organic fertilizers with antibiotic residues, resistance genes, and mobile genetic elements mirroring antibiotic use in livestock? Applied microbiology and biotechnology. 2016;100(21):9343-53.
5. Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J, Handelsman J. Call of the wild: antibiotic resistance genes in natural environments. Nature reviews microbiology. 2010;8(4):251.
6. Wellington EM, Boxall AB, Cross P, Feil EJ, Gaze WH, Hawkey PM, et al. The role of the natural environment in the emergence of antibiotic resistance in Gram-negative bacteria. The Lancet infectious diseases. 2013;13(2):155-65.

7. Zhu Y-G, Zhao Y, Li B, Huang C-L, Zhang S-Y, Yu S, et al. Continental-scale pollution of estuaries with antibiotic resistance genes. Nature microbiology. 2017;2(4):16270.
8. Brandt C, Braun SD, Stein C, Slickers P, Ehricht R, Pletz MW, et al. In silico serine β-lactamases analysis reveals a huge potential resistome in environmental and pathogenic species. Scientific reports. 2017;7:43232.
9. Ambler R, Coulson A, Frere J-M, Ghuysen J-M, Joris B, Forsman M, et al. A standard numbering scheme for the class A beta-lactamases. Biochemical Journal. 1991;276(Pt 1):269.
10. Bush K, Jacoby GA, Medeiros AA. A functional classification scheme for beta-lactamases and its correlation with molecular structure. Antimicrobial agents and chemotherapy. 1995;39(6):1211.
11. Bush K, Jacoby GA. Updated functional classification of β-lactamases. Antimicrobial agents and chemotherapy. 2010;54(3):969-76.
12. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 2018;6(1):1-15.
13. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12(2):85-94; doi: 10.1093/protein/12.2.85.
14. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. Journal of antimicrobial chemotherapy. 2012;67(11):2640-4.
15. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic acids research. 2016:gkw1004.
16. Liu B, Pop M. ARDB—antibiotic resistance genes database. Nucleic acids research. 2008;37(suppl_1):D443-D7.
17. Yang Y, Jiang X, Chai B, Ma L, Li B, Zhang A, et al. ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. Bioinformatics. 2016;32(15):2346-51.
18. Yin X, Jiang X-T, Chai B, Li L, Yang Y, Cole JR, et al. ARGs-OAP v2. 0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. Bioinformatics. 2018;34(13):2263-70.
19. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 2018;6(1):23.
20. Berglund F, Österlund T, Boulund F, Marathe NP, Larsson DJ, Kristiansson E. Identification and reconstruction of novel antibiotic resistance genes from metagenomes. Microbiome. 2019;7(1):52.
21. Monstein H-J, Tärnberg M, Nilsson LE. Molecular identification of CTX-M and bla OXY/K1 β-lactamase genes in Enterobacteriaceae by sequencing of universal M13-sequence tagged PCR-amplicons. BMC infectious diseases. 2009;9(1):7.

22. Tärnberg M, Nilsson LE, Monstein H-J. Molecular identification of blaSHV, blaLEN and blaOKP β-lactamase genes in Klebsiella pneumoniae by bi-directional sequencing of universal SP6-and T7-sequence-tagged blaSHV-PCR amplicons. Molecular and cellular probes. 2009;23(3-4):195-200.
23. McArthur AG, Tsang KK. Antimicrobial resistance surveillance in the genomic age. Annals of the New York Academy of Sciences. 2017;1388(1):78-91.
24. Young AL, Nicol MP, Moodley C, Bamford CM. The accuracy of extended-spectrum beta-lactamase detection in Escherichia coli and Klebsiella pneumoniae in South African laboratories using the Vitek 2 Gram-negative susceptibility card AST-N255. Southern African Journal of Infectious Diseases. 2019;34(1):1-6.
25. Orellana LH, Rodriguez-R LM, Konstantinidis KT. ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. Nucleic acids research. 2016;45(3):e14-e.
26. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic acids research. 2012;40(12):e94-e.
27. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011;7(1).
28. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC bioinformatics. 2009;10(1):421.
29. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015;12(1):59.
30. Thai QK, Bös F, Pleiss J. The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. BMC genomics. 2009;10(1):390.
31. Fischer M, Thai QK, Grieb M, Pleiss J. DWARF—a data warehouse system for analyzing protein families. BMC bioinformatics. 2006;7(1):495.
32. Buchholz PC, Vogel C, Reusch W, Pohl M, Rother D, Spieß AC, et al. BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments. ChemBioChem. 2016;17(21):2093-8.
33. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-2.
34. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9.
35. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.
36. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.
37. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic acids research. 2013;41(12):e121-e.

38. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, et al. FunGene: the functional gene pipeline and repository. Frontiers in Microbiology. 2013;4:291.
39. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic acids research. 2018;47(D1):D427-D32.
40. Rice P, Longden I, Bleasby A: EMBOSS: the European molecular biology open software suite. In.: Elsevier current trends; 2000.
41. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research. 2016;44(W1):W242-W5.
42. Rodriguez-R LM, Konstantinidis KT: The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. In.: PeerJ Preprints; 2016.
43. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. PLoS computational biology. 2015;11(12):e1004557; doi: 10.1371/journal.pcbi.1004557.
44. ALTSCHUL SF, MADDEN TL, SCHÄFFER AA, ZHANG J, ZHANG Z, MILLER W, et al. A new generation of protein database search programs. Nucleic Acids Research. 25(17).
45. Prestat E, David MM, Hultman J, Taş N, Lamendella R, Dvornik J, et al. FOAM (functional ontology assignments for metagenomes): a hidden Markov model (HMM) database with environmental focus. Nucleic acids research. 2014;42(19):e145-e.
46. Mande SS, Mohammed MH, Ghosh TS. Classification of metagenomic sequences: methods and challenges. Briefings in bioinformatics. 2012;13(6):669-81.
47. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome research. 2007;17(3):377-86.
48. Gerlach W, Stoye J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic acids research. 2011;39(14):e91-e.
49. Meini M-R, Llarrull LI, Vila AJ. Evolution of metallo-β-lactamases: trends revealed by natural diversity and in vitro evolution. Antibiotics. 2014;3(3):285-316.
50. Haeggman S, Löfdahl S, Paauw A, Verhoef J, Brisse S. Diversity and evolution of the class A chromosomal beta-lactamase gene in Klebsiella pneumoniae. Antimicrobial agents and chemotherapy. 2004;48(7):2400-8.
51. Fevre C, Jbel M, Passet V, Weill F-X, Grimont PA, Brisse S. Six groups of the OXY β-lactamase evolved over millions of years in Klebsiella oxytoca. Antimicrobial agents and chemotherapy. 2005;49(8):3453-62.
52. Lanza VF, Baquero F, Martinez JL, Ramos-Ruiz R, Gonzalez-Zorn B, Andremont A, et al. In-depth resistome analysis by targeted metagenomics. Microbiome. 2018;6(1):11; doi: 10.1186/s40168-017-0387-y.
53. Taitt CR, Leski TA, Stockelman MG, Craft DW, Zurawski DV, Kirkup BC, et al. Antimicrobial resistance determinants in Acinetobacter baumannii isolates taken from military treatment facilities. Antimicrob

Agents Chemother. 2014;58(2):767-81; doi: 10.1128/AAC.01897-13.

54. Van Boeckel TP, Gandra S, Ashok A, Caudron Q, Grenfell BT, Levin SA, et al. Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. The Lancet Infectious Diseases. 2014;14(8):742-50.

Figures

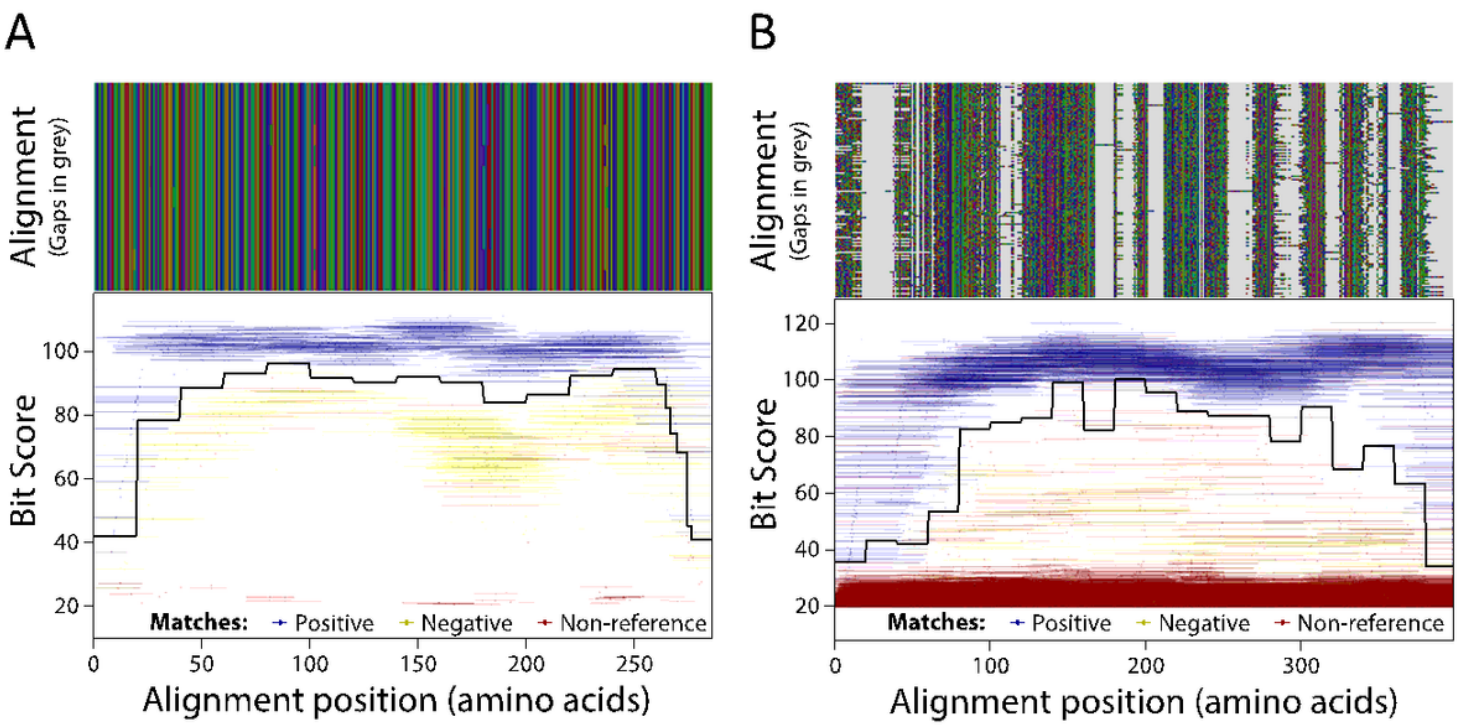

Figure 1

Plots of the 150 bp ROCker model of (A) TEM using a DIAMOND search and (B) the class D BL after trimming of the 5' conserved end of the alignment. The top panels display the sequence alignments with amino acids encoded in different colors and gaps in light grey. The middle panels represent the bit score (y-axis) of the reads of the 150 bp training data set against the positive reference sequences used to build the ROCker model carrying positive reference (blue), negative reference (yellow), or non-target sequences (red). Each matching read is represented as a line based on the coordinates of the alignment of the positive reference sequences to which the read maps, and dots represent the midpoint of each read. The solid black traversing line represents the calculated ROCker best bit score thresholds for consecutive windows of variable length. The sensitivity, specificity and accuracy were 98.8%, 99.2% 99.0%, and 99.1%, 99.8%, 99.8% for the TEM and class D BL model, respectively.

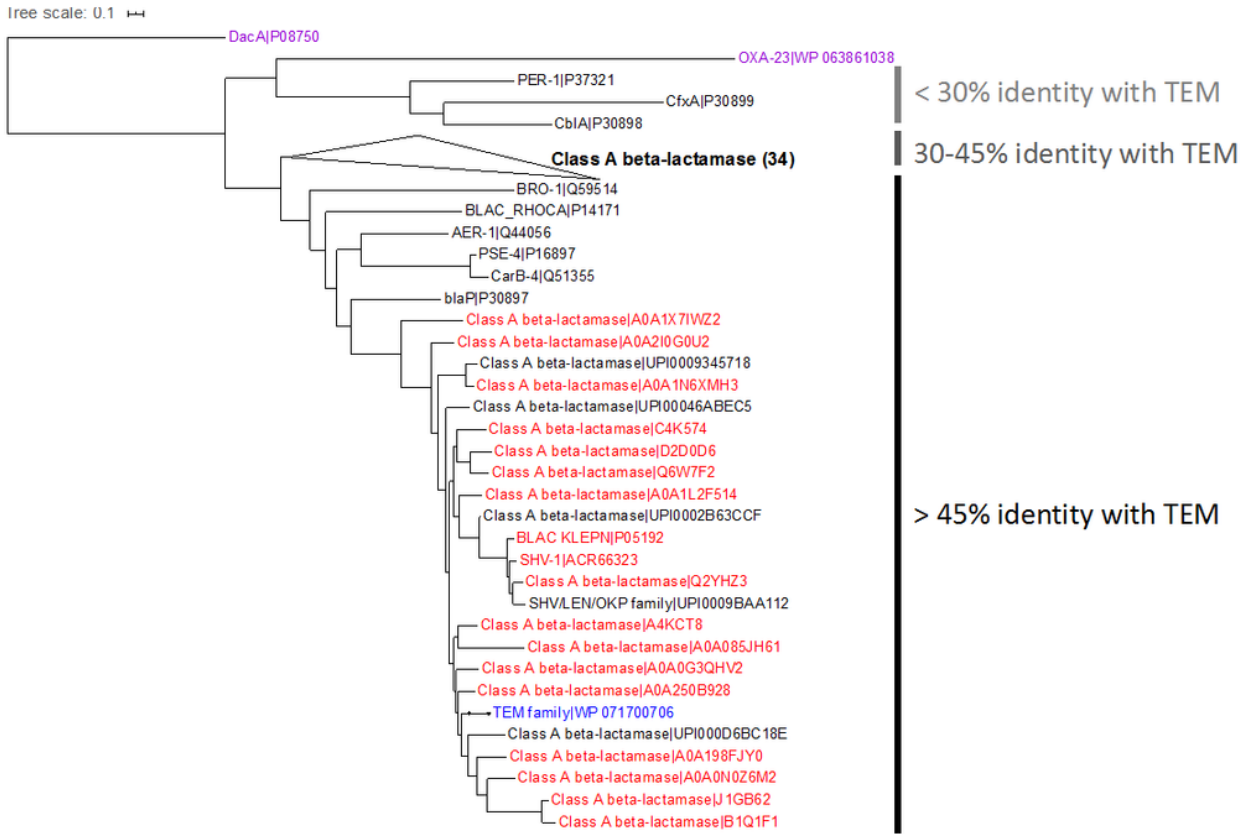

Figure 2

Phylogenetic relationships of TEM (in blue) with other families of the class A β-lactamases. D-alanyl-D-alanine carboxypeptidase (DacA, outgroup) and OXA-23 (class D β-lactamase) were included for comparison (in purple, at the top of the figure). The references were first clustered by 90% amino acid identity using cd-hit and only one representative per resulting cluster was included. The phylogenetic tree was constructed using maximum likelihood in RAxML version 7.7.2 with the GTRGAMMA model. The references in red (beta-lactamases that are not of the TEM family) were included as negative references used in building the ROCker model. See Materials and Methods for details on reference sequence selection.

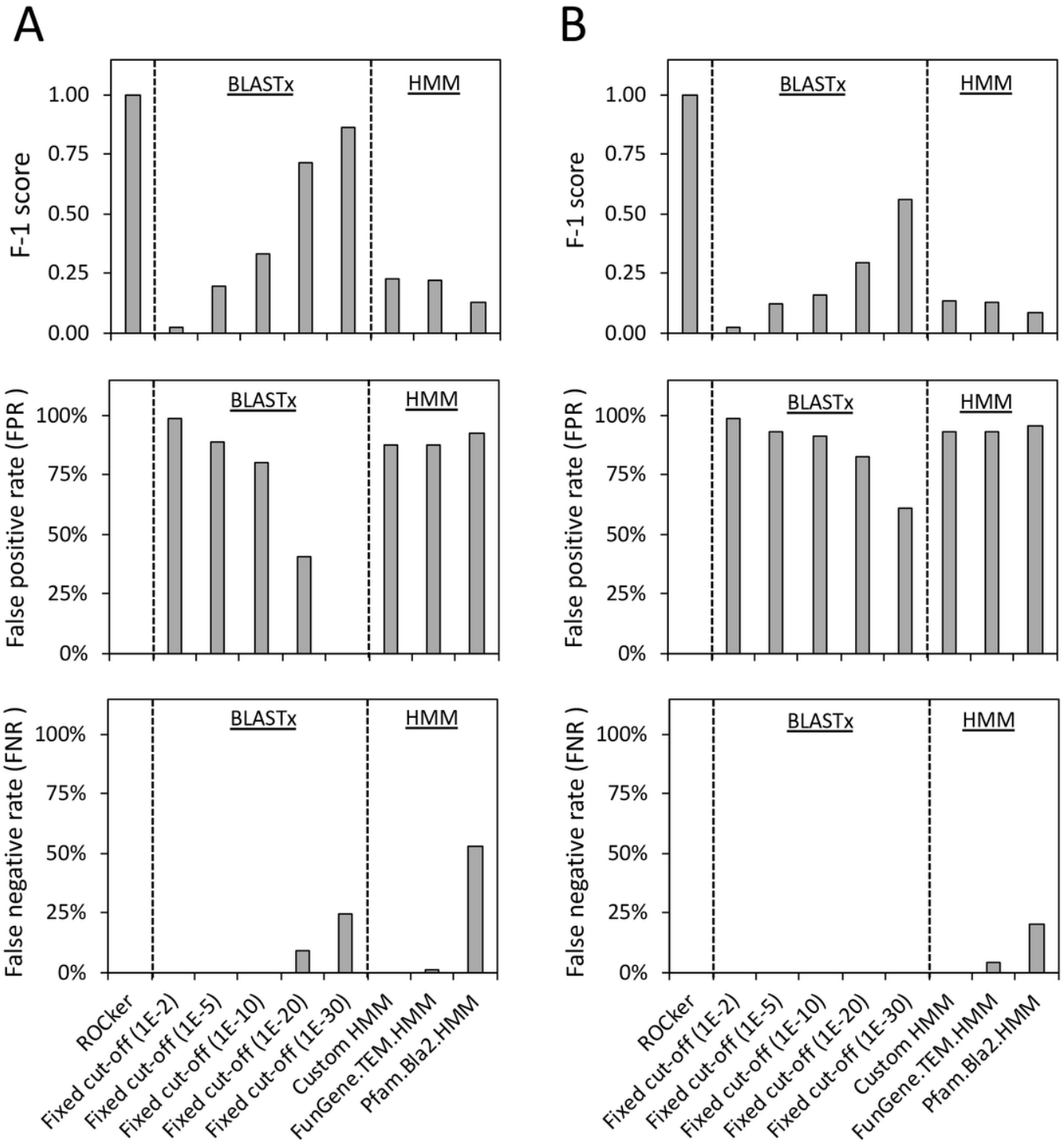

Figure 3

Performance of the TEM ROCker model on 150 bp (A) and 250 bp (B) mock data sets compared to alternative methods. Alternative methods included BLASTx searches using a fixed e-value (10-2, 10-5, 10-10, 10-20 and 10-30) and HMM-based searches using HMMs (e-value < 0.1) downloaded from FunGene or Pfam databases, or a custom-built HMM using the same positive reference sequence set used to build the ROCker model (x-axis labels). F1-score [2*precision*recall/(precision+recall)] (top), false positive rate [FPR = FP/(TP+FP)] (middle), and false negative rate [FNR = FN/(TP+FN)] (bottom) statistics are shown.

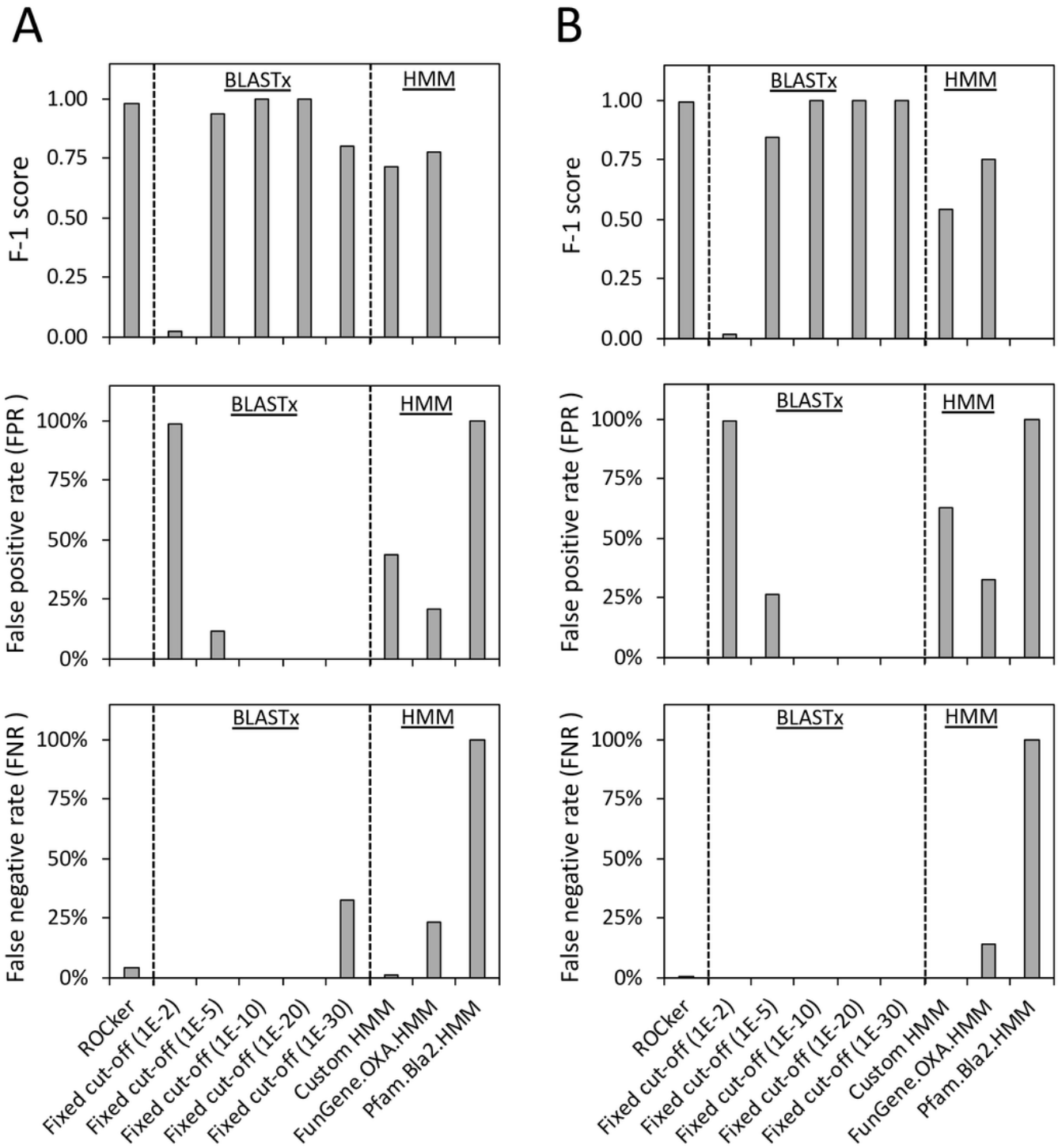

Figure 4

Performance of the class D BL ROCker model on 150 bp (A) and 250 bp (B) mock data sets compared to alternative methods. The same methods and parameters were used as described for Figure 3. F1-score [2*precision*recall/(precision+recall)] (top), false positive rate [FPR = FP/(TP+FP)] (middle), and false negative rate [FNR = FN/(TP+FN)] (bottom) statistics are shown.

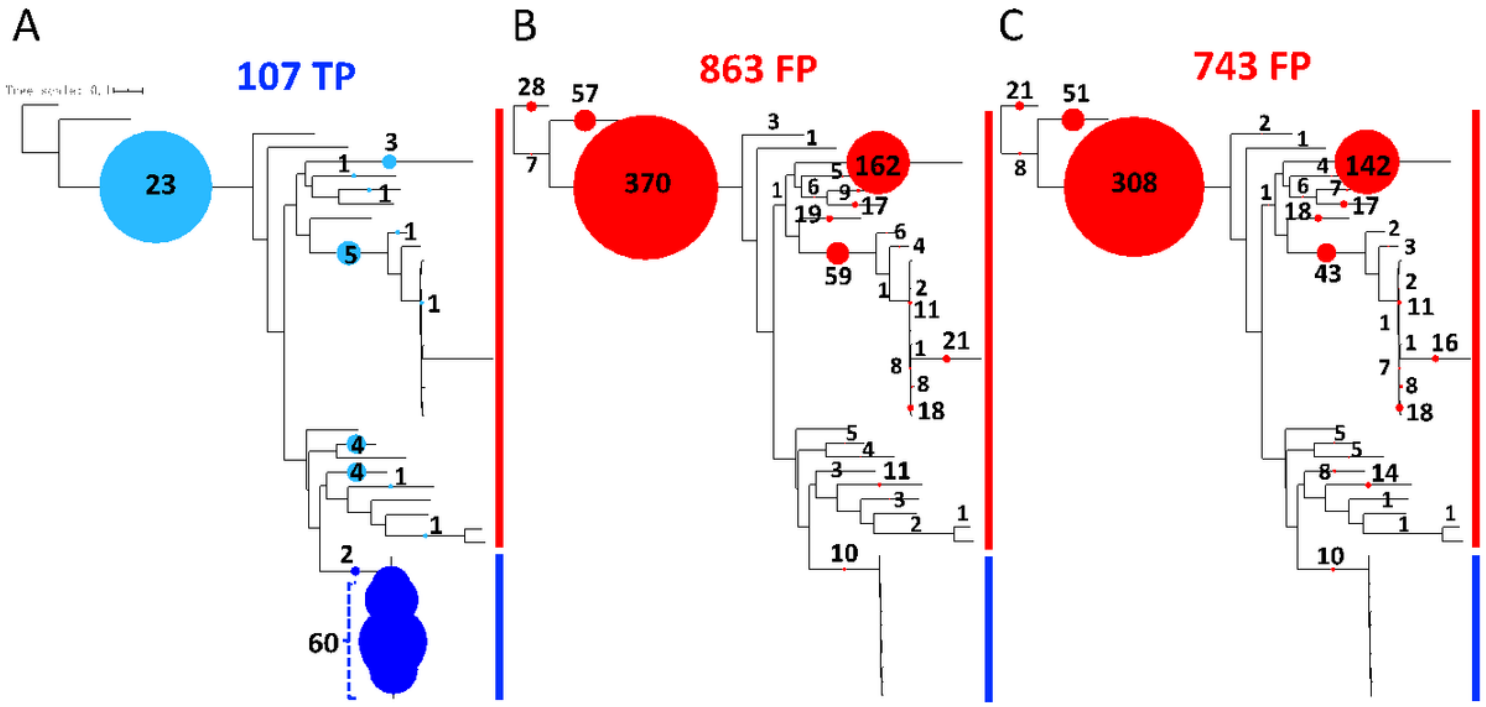

Figure 5

Phylogenetic placement of the blaTEM-carrying (true positive) and false positive reads (from non-target genes) identified in the 150 bp mock data set by different approaches. The bar to the right of each tree emphasizes clades made up of target (blue) and non-target (red) reference proteins. The colored circles represent the number of reads that were placed on the corresponding branches. The number of reads that make up each circle are provided for direct comparison between the trees (the filled circles on the same tree are proportional to the number of reads and directly comparable to each other, but not directly comparable between trees). The positive references used to build the TEM ROCker model are colored in blue (reads originating from these sequences are target reads and are included in the blue circles as TP) and negative references used to build TEM ROCker model are displayed in red (reads originating from these sequences are non-target reads and are included in the red circles as FP). (A) Placement of 107 true positive reads identified in common by ROCker, BLASTx search (e-value < 10-5) and searches with a custom-built TEM HMM (e-value < 0.1). Note that several of these reads (in light blue) were (erroneously) placed to non-target clades due to lack of phylogenetic signal (e.g., the reads sampled highly conserved regions between target and non-target references proteins) or sequence errors that confused the placement pipeline. The former reads were usually placed on ancestral nodes while the latter to the terminal nodes. (B) Placement of 863 false positive reads identified by BLASTx search with e-value < 10-5. (C) Placement of 743 false positive reads identified by searches with a custom-built TEM HMM with e-value < 0.1. Note that a small number of these FP reads (n=10) were (erroneously) placed on the clade made up of target sequences due to inadequate phylogenetic signal. The phylogenetic tree was constructed using maximum likelihood in RAxML version 7.7.2 with the GTRGAMMA model. Reads were

placed on the phylogenetic tree as described in the Materials and Methods section. Note that no false negatives (i.e., blaTEM-carrying reads that were not detected by the tools) were observed by ROCker, BLASTx search (e-value < 10-5) and searches with a custom-built TEM HMM (e-value < 0.1) for this BL family.

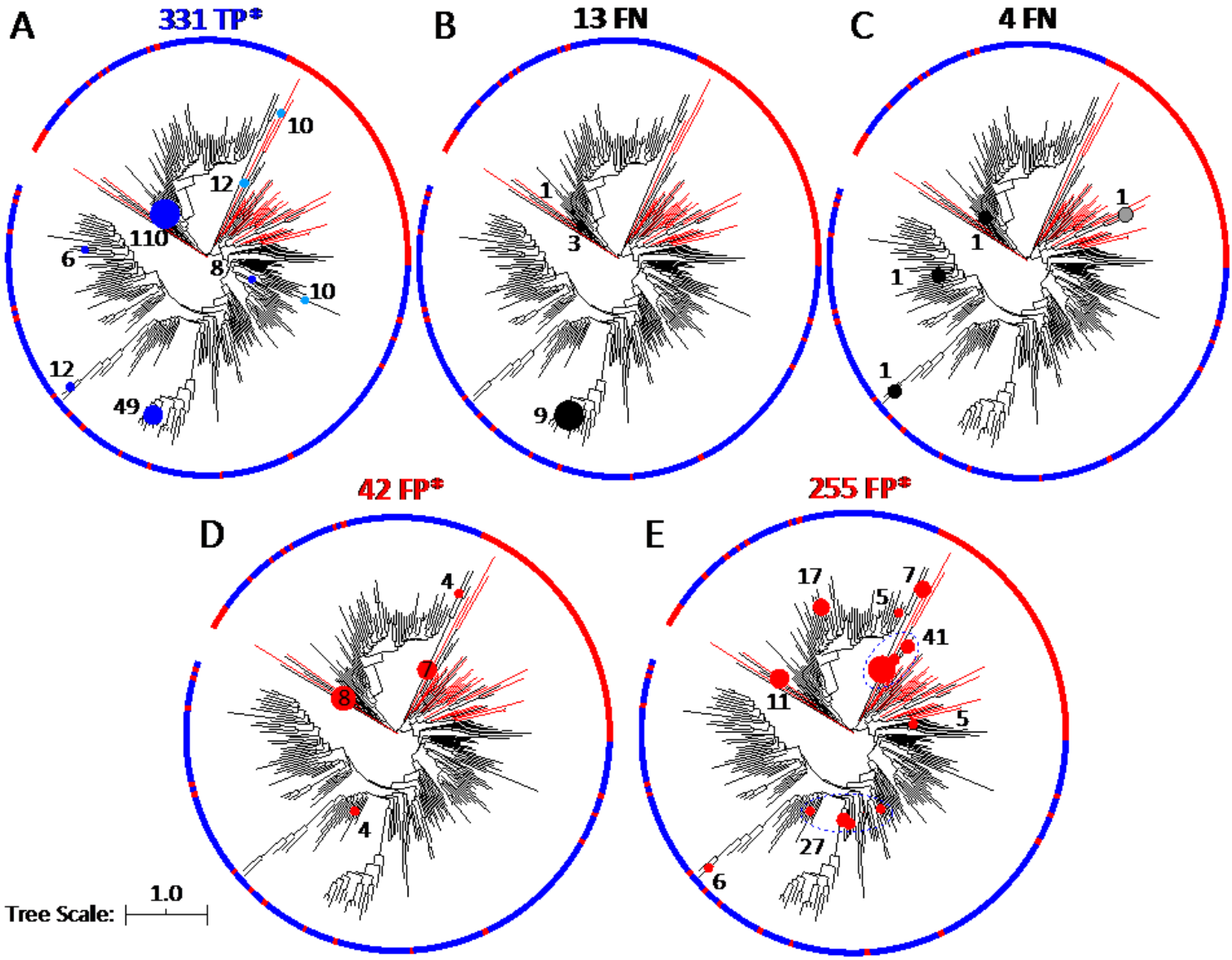

Figure 6

Phylogenetic placement of the class D BL gene-carrying reads (true positives and false negatives) and reads from non-target genes (false positive reads) identified in the 150 bp mock data set using different approaches. The same methods and parameters were used as described for Figure 5 but a circular tree was chosen due to the higher number and more diverse sequences of the class D BLs, and in panels A, D, and E (denoted by an asterisk), not all identified reads are shown for demonstration purposes (e.g., many clades recruited only a couple reads). In addition, red branches indicate UniRef90 proteins that were included in the negative reference set for ROCker model building (see also a more detailed class D BL phylogeny in Supplementary Figure S1). (A) True positive reads (n=331; 114 reads in 58 clades are not shown) detected by BLASTx search (e-value < 10-5). (B) False negative reads (n=13) missed by ROCker

(i.e. reads carrying class D BLs that ROCker did not detect). (C) False negative reads (n=4) missed by the custom-built class D BL HMM. (D) False positive reads (n=42; 19 reads in 15 clades are not shown) detected by BLASTx search (e-value < 10-5). (E) False positive reads (n=255; 135 reads in 83 clades are not shown) detected by the custom-built class D BL HMM. As in Figure 5, circles in light blue or gray denote reads that were placed onto non-target genes or clades due to inadequate phylogenetic signal.

Supplementary Files

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

- BLROCkerModelSupplementaryTable.xlsx
- BLROCkerModelSUPPLEMENTARY.docx