Sequencing-based methods and resources to study antimicrobial resistance

Manish Boolchandani1,5, Alaric W. D’Souza1,5 and Gautam Dantas1,2,3,4*

Abstract | Antimicrobial resistance extracts high morbidity, mortality and economic costs yearly by rendering bacteria immune to antibiotics. Identifying and understanding antimicrobial resistance are imperative for clinical practice to treat resistant infections and for public health efforts to limit the spread of resistance. Technologies such as next-generation sequencing are expanding our abilities to detect and study antimicrobial resistance. This Review provides a detailed overview of antimicrobial resistance identification and characterization methods, from traditional antimicrobial susceptibility testing to recent deep-learning methods. We focus on sequencing-based resistance discovery and discuss tools and databases used in antimicrobial resistance studies.

Antimicrobials are small molecules that can inhibit or kill bacteria. These small molecules are commonly used as therapeutics for bacterial infections, but some bacteria can grow and survive despite antimicrobial pressures, a property known as antimicrobial resistance. In clinical settings, resistant bacterial infections decrease available treatment options and increase morbidity and mortality compared with those caused by susceptible bacteria. Resistance is observed against nearly all antimicrobials (Fig. 1a,b), including so-called last-resort antimicrobials used in life-threatening, multidrug-resistant infections. Bacteria resistant to first-line antimicrobials infect 2 million people in the USA yearly, and these infections exact a US$20 billion health-care cost. This problem is not isolated to the USA. In the European Union, antimicrobial resistance has accounted for >30,000 deaths and nearly 900,000 disability-adjusted life-years. In fact, multiple national and global public health organizations categorize antimicrobial resistance as an imminent danger and uniformly agree that tracking its emergence and prevalence is critical to minimize the threat to human health.

The traditional method for assessing antimicrobial resistance in bacteria is antimicrobial susceptibility testing (AST). AST is viable only for cultivable bacteria, precluding studies on the emergence and spread of antimicrobial resistance in diverse and complex microbial communities with large fractions of currently uncultured bacteria. Bacterial antimicrobial resistance is usually genetically encoded. Genetically encoded antimicrobial resistance can occur through a number of mechanisms, including overexpression or duplication of existing genes, point mutations or the acquisition of entirely new genes via horizontal gene transfer (HGT). Improvements in next-generation sequencing technologies and computational methods are facilitating rapid antimicrobial resistance gene identification and characterization in genomes and metagenomes. These developing technologies and methods complement traditional culture-based methods for clinical and surveillance applications and provide opportunities for quick and sensitive resistance determinations in cultivable and uncultivable bacteria. Large-scale and comparative studies of human, animal and environmental samples have provided unprecedented insights into the global distribution of antimicrobial resistance genes and the spread of multidrug-resistant bacteria, resistance exchange networks and how different habitats and phylogeny affect the evolutionary dynamics of antimicrobial resistance worldwide. Understanding and surveying genetic determinants of resistance using sequencing data pose unique challenges that are being addressed by the tools and databases reviewed here.
Antibiotic targets in bacterial cells

- Cell wall
  - Glycopeptides (vancomycin)
  - β-lactams (penicillin)
  - Ethambutol (ethambutol)
  - Isoniazid (isoniazid)
  - Fosfomycin (fosfomycin)
  - Bacitracin (bacitracin)
  - Cycloserine (cycloserine*)

- Cell membrane
  - Macrolides (azithromycin)
  - Streptogramins (dalfopristin)
  - Amphenicol (chloramphenicol)
  - Lincosamide (clindamycin)
  - Oxazolidinone (linezolid)
  - Fusidanes (fusidic acid)

- Nucleic acid synthesis
  - DNA gyrase
  - Ribosome

- mRNA transcription

- DNA synthesis

- Nucleic acid modification

- Ribosome

- RNA synthetase

- Target protection

- Antibiotic efflux

- Reduced permeability

- Antibiotic modification/degradation

- Expression changes

- Target modification (includes acquisition of insensitive functional targets)

Mechanisms of antibiotic resistance

- Susceptible
- Resistant

- Antibiotic
- Target protein

Genetic determinants of resistance

- Horizontal gene transfer
- Gene duplication
- New promoter
- Point mutation
- Insertion or deletion

Resistance or target gene
Promoter

Fig. 1 | Antimicrobial targets and resistance mechanisms. a | Antimicrobials are grouped by target site. Drug classes are in bold, and example drugs from that class are in parentheses below the drug classes. Antimicrobial resistance mechanisms that act on that class are depicted left of the antimicrobial class according to the layout shown in part b. b | Mechanisms of antimicrobial resistance are depicted with susceptible organisms represented on the left and resistant organisms represented on the right. To the left of each labelled mechanism is the legend annotation position used in part a. c | Genetic underpinnings of antimicrobial resistance are illustrated.
Bacterial culture has a long history as an integral part of clinical microbiology. Researchers cultivate bacteria on agar plates or in liquid broth to probe bacterial phenotypes and to discover novel bacterial functions. Hospital clinical microbiology laboratories can use data gained in these assays to inform clinical treatment decisions.

**Current techniques**

For phenotypic testing, bacteria are isolated from patient or environmental samples by non-selective or selective agar plating for pure colonies. Isolated bacteria are then directly challenged with antimicrobials to test for antimicrobial resistance. In liquid media microbroth dilution antimicrobial susceptibility testing (AST), bacteria are challenged with decreasing antimicrobial concentrations to find the concentration at which they grow successfully. The lowest antibiotic concentration that inhibits the bacterial growth is called the minimum inhibitory concentration. Solid media techniques use antimicrobial Kirby–Bauer disc diffusion or gradient diffusion strips to measure the clearance of the bacteria from the source of the antimicrobial. The area around the antimicrobial disc or strip is called the zone of inhibition (culture-based AST) and can take weeks.

This rapid testing is especially valuable for slow-growing organisms such as Mycobacterium tuberculosis, for which culture techniques can also fail in situations where multiple bacteria may cause symptomatic disease. In such cases, the cultured bacteria are assumed responsible, but several metagenomic studies from patient and metagenomic samples. We also describe recent advancements to mitigate weaknesses in resistance detection methods, and we highlight areas requiring greater focus.

**Challenges and remaining gaps**

While AST is useful to provide phenotypic resistance results, it is low throughput. For each sample, a clinical microbiologist needs to culture the bacteria and to set up and read the AST results. This is limiting because not all health-care centres have trained clinical microbiologists on staff. For some bacteria (for example, M. tuberculosis), current laboratory diagnostic techniques feasible in low-resource areas have low sensitivity. Susceptibility standards are also not published for all combinations of antimicrobials and infection-causing, culturable bacteria. Additionally, the standards themselves are not uniform across different countries. Culture techniques can also fail in situations where multiple bacteria may cause symptomatic disease. In such cases, the cultured bacteria are assumed responsible, but several metagenomic studies from patient samples indicate that the bacteria detected in culture may not be responsible for disease symptoms. Finally, these phenotypic methods have lower resolution in examining resistance determinants and provide less information about resistance gene epidemiology than whole-genome sequencing.

**Innovations**

Automation is rapidly being developed and implemented to conduct phenotypic assays, including AST. Several systems, including VITEK, ADAGIO, Accelerate Pheno and others, are already moving into clinical spaces. The automated platforms have several key advantages, including continuous monitoring systems, sensitive optical instruments for measuring susceptibility results and standardized internal clocks. These advantages can speed up susceptibility results for cultured bacteria. They may also increase consistency of susceptibility results across different locations and reduce the burden of work for clinical microbiologists. Unfortunately, these automated platforms can be prohibitively expensive to set up.

**Sequencing-based resistance discovery**

Advancements in sequencing technologies have increased bacterial sequence data availability, and continually decreasing costs have made sequencing a viable antimicrobial resistance surveillance tool. Several methods and tools have been published in recent years for detecting genetic determinants of antimicrobial resistance from whole-genome sequencing (WGS) and whole-metagenome sequencing (WMS) datasets. Organizing sequencing data is an important pre-processing step before antimicrobial resistance gene analysis. Short reads, generated by technologies such as Illumina, can either be processed using assembly-based methods, whereby sequencing reads are first assembled into contiguous fragments and then annotated by comparing with custom or public reference databases, or directly analysed using read-based methods, whereby resistance determinants are predicted by mapping reads directly to a reference database.

**Assembly-based methods.** The de novo assembly of WGS of bacterial genomes from short-read data is generally performed by De Bruijn graph (DBG)-based assemblers such as SPAdes, Velvet and SOAPdenovo. In this approach, sequencing reads are divided into shorter overlapping sub-sequences (called k-mers) of length k (where k is less than the read length) and are used to form a network graph. The assemblers then reconstruct the genome sequence by finding an optimum path (Euler’s path) through the graph that visits each edge once. Errors in sequencing data introduce false k-mers in the graph, resulting in fragmented assemblies. Several assemblers (for example, SPAdes and Velvet) heuristically eliminate these errors before finding an Euler’s path in the graph. Assembling WMS data is more complicated than single-isolate assembly, as the algorithms need to account for unknown abundances of different organisms with unknown phylogenetic relationships. In single-genome assembly, uniform sequencing coverage across the genome is used by assemblers to correct sequencing errors and to identify repetitive sequences and plasmids. Several assemblers exploit the higher coverage of plasmids owing to copy number to distinguish between chromosome and plasmid sequences in isolate genomes but uneven coverage of different organisms in WMS data makes detecting repeats difficult. Long stretches of identical sequences in unrelated species further complicate assembly by making it difficult to assign reads to a particular species. Thus, algorithms developed for single-genome assembly cannot be directly applied to assemble metagenomes. Several metagenome-specific assemblers have been developed to overcome these challenges, either by partitioning or optimizing the graph for uneven sequencing depths. Some notable metagenome assemblers are IDBA-UD, MEGAHIT, MetaSPAdes and MetaVelvet (extensions of SPAdes and Velvet for metagenomes). The CAMI project.

**Antimicrobial susceptibility testing (AST)**

Challenge of bacteria with antimicrobials to determine whether they have phenotypic antimicrobial resistance.

---

**Box 1 | Culture-based susceptibility testing**

Bacterial culture has a long history as an integral part of clinical microbiology. Researchers cultivate bacteria on agar plates or in liquid broth to probe bacterial phenotypes and to discover novel bacterial functions. Hospital clinical microbiology laboratories can use data gained in these assays to inform clinical treatment decisions.

For phenotypic testing, bacteria are isolated from patient or environmental samples by non-selective or selective agar plating for pure colonies. Isolated bacteria are then directly challenged with antimicrobials to test for antimicrobial resistance. In liquid media microbroth dilution antimicrobial susceptibility testing (AST), bacteria are challenged with decreasing antimicrobial concentrations to find the concentration at which they grow successfully. The lowest antibiotic concentration that inhibits the bacterial growth is called the minimum inhibitory concentration. Solid media techniques use antimicrobial Kirby–Bauer disc diffusion or gradient diffusion strips to measure the clearance of the bacteria from the source of the antimicrobial. The area around the antimicrobial disc or strip is called the zone of inhibition (culture-based AST), for which culture-based AST can take weeks.

This rapid testing is especially valuable for slow-growing organisms such as Mycobacterium tuberculosis, for which culture techniques can also fail in situations where multiple bacteria may cause symptomatic disease. In such cases, the cultured bacteria are assumed responsible, but several metagenomic studies from patient and metagenomic samples. We also describe recent advancements to mitigate weaknesses in resistance detection methods, and we highlight areas requiring greater focus.

---

**Sequencing-based resistance discovery**

Advancements in sequencing technologies have increased bacterial sequence data availability, and continually decreasing costs have made sequencing a viable antimicrobial resistance surveillance tool. Several methods and tools have been published in recent years for detecting genetic determinants of antimicrobial resistance from whole-genome sequencing (WGS) and whole-metagenome sequencing (WMS) datasets. Organizing sequencing data is an important pre-processing step before antimicrobial resistance gene analysis. Short reads, generated by technologies such as Illumina, can either be processed using assembly-based methods, whereby sequencing reads are first assembled into contiguous fragments (contigs) and then annotated by comparing with custom or public reference databases, or directly analysed using read-based methods, whereby resistance determinants are predicted by mapping reads directly to a reference database.

**Assembly-based methods.** The de novo assembly of WGS of bacterial genomes from short-read data is generally performed by De Bruijn graph (DBG)-based assemblers such as SPAdes, Velvet and SOAPdenovo. In this approach, sequencing reads are divided into shorter overlapping sub-sequences (called k-mers) of length k (where k is less than the read length) and are used to form a network graph. The assemblers then reconstruct the genome sequence by finding an optimum path (Euler’s path) through the graph that visits each edge once. Errors in sequencing data introduce false k-mers in the graph, resulting in fragmented assemblies. Several assemblers (for example, SPAdes and Velvet) heuristically eliminate these errors before finding an Euler’s path in the graph. Assembling WMS data is more complicated than single-isolate assembly, as the algorithms need to account for unknown abundances of different organisms with unknown phylogenetic relationships. In single-genome assembly, uniform sequencing coverage across the genome is used by assemblers to correct sequencing errors and to identify repetitive sequences and plasmids — several assemblers exploit the higher coverage of plasmids owing to copy number to distinguish between chromosome and plasmid sequences in isolate genomes but uneven coverage of different organisms in WMS data makes detecting repeats difficult. Long stretches of identical sequences in unrelated species further complicate assembly by making it difficult to assign reads to a particular species. Thus, algorithms developed for single-genome assembly cannot be directly applied to assemble metagenomes. Several metagenome-specific assemblers have been developed to overcome these challenges, either by partitioning or optimizing the graph for uneven sequencing depths. Some notable metagenome assemblers are IDBA-UD, MEGAHIT, MetaSPAdes and MetaVelvet (extensions of SPAdes and Velvet for metagenomes). The CAMI project.
**Table 1 | Sequencing-based tools for antimicrobial resistance detection**

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Accessibility</th>
<th>Year</th>
<th>Link</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assembly-based tools</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resfinder⁷²</td>
<td>Tool for detecting acquired AR genes from sequenced or partially sequenced bacterial isolates</td>
<td>Web and/or standalone</td>
<td>2012</td>
<td><a href="https://cge.cbs.dtu.dk/services/ResFinder/">https://cge.cbs.dtu.dk/services/ResFinder/</a></td>
<td>Active</td>
</tr>
<tr>
<td>ARG-ANNOT⁶⁶</td>
<td>Tool for pairwise comparison of query sequence with ARG-ANNOT database</td>
<td>Standalone</td>
<td>2014</td>
<td>Not available</td>
<td>Archived</td>
</tr>
<tr>
<td>RGI⁸⁶</td>
<td>• Pairwise comparison of query sequence with the CARD</td>
<td>Web and/or standalone</td>
<td>2015</td>
<td><a href="https://card.mcmaster.ca/analyze/rgi">https://card.mcmaster.ca/analyze/rgi</a></td>
<td>Active</td>
</tr>
<tr>
<td></td>
<td>• Uses curated AR detection models to predict intrinsic resistance genes, dedicated resistance genes and acquired resistance from mutations in drug targets</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARGs-OAP (v2)⁴¹</td>
<td>• Online analysis pipeline for AR genes</td>
<td>Web and/or standalone</td>
<td>2016</td>
<td><a href="https://galaxyproject.org/use/args-oap/">https://galaxyproject.org/use/args-oap/</a></td>
<td>Active</td>
</tr>
<tr>
<td></td>
<td>• Detection from metagenomic data using an integrated structured database of AR sequences</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARIBA⁵²</td>
<td>Tool for rapid AR genotyping directly from sequencing reads using curated public databases</td>
<td>Standalone</td>
<td>2017</td>
<td><a href="https://github.com/sanger-pathogens/ARIBA">https://github.com/sanger-pathogens/ARIBA</a></td>
<td>Active</td>
</tr>
<tr>
<td>PointFinder⁷¹</td>
<td>Web tool for WGS-based detection of AR associated with chromosomal point mutations in bacterial pathogens</td>
<td>Web and/or standalone</td>
<td>2018</td>
<td><a href="https://cge.cbs.dtu.dk/services/PointFinder/">https://cge.cbs.dtu.dk/services/PointFinder/</a></td>
<td>Active</td>
</tr>
<tr>
<td><strong>Read-based tools</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRST2 [REF ⁵⁹]</td>
<td>Tool for direct mapping of reads to curated AR databases</td>
<td>Standalone</td>
<td>2014</td>
<td><a href="https://github.com/katholt/srst2">https://github.com/katholt/srst2</a></td>
<td>Active</td>
</tr>
<tr>
<td>SEAR⁴¹</td>
<td>Cloud-compatible pipeline and web interface for rapidly detecting AR genes directly from sequencing reads</td>
<td>Web and/or standalone (archived)</td>
<td>2015</td>
<td><a href="https://github.com/will-rowe/SEAR">https://github.com/will-rowe/SEAR</a></td>
<td>Archived</td>
</tr>
<tr>
<td>ShortBRED⁶¹</td>
<td>Tool to profile protein families in the metagenomic data using short peptide marker sequences</td>
<td>Standalone</td>
<td>2015</td>
<td><a href="http://butterhown.sph.harvard.edu/shortbred">http://butterhown.sph.harvard.edu/shortbred</a></td>
<td>Active</td>
</tr>
<tr>
<td>PATRIC⁷⁶</td>
<td>A unique resource for studying AR</td>
<td>Web</td>
<td>2016</td>
<td><a href="http://www.paticbrc.org">www.paticbrc.org</a></td>
<td>Active</td>
</tr>
<tr>
<td>SSTAR⁷⁷</td>
<td>Tool to identify known, putative new alleles and truncated versions of existing AR genes from WGS data</td>
<td>Standalone</td>
<td>2016</td>
<td><a href="https://github.com/tomdeman-bio/Sequence-Search-Tool-for-Antimicrobial-Resistance-SSTAR/">https://github.com/tomdeman-bio/Sequence-Search-Tool-for-Antimicrobial-Resistance-SSTAR/</a></td>
<td>Active</td>
</tr>
<tr>
<td>KmerResistance⁵⁵</td>
<td>Program that compares the co-occurrence of k-mers from raw reads with k-mers from multiple databases</td>
<td>Web</td>
<td>2016</td>
<td><a href="https://cge.cbs.dtu.dk/services/KmerResistance/">https://cge.cbs.dtu.dk/services/KmerResistance/</a></td>
<td>Active</td>
</tr>
<tr>
<td>GROOT⁵⁶</td>
<td>Software that enables resistome profiling by mapping metagenomic reads to graph representation of reference gene sets</td>
<td>Standalone</td>
<td>2018</td>
<td><a href="https://github.com/will-rowe/groot">https://github.com/will-rowe/groot</a></td>
<td>Active</td>
</tr>
<tr>
<td>DeepArgs⁹⁰</td>
<td>A deep-learning approach for predicting AR genes from metagenomic data</td>
<td>Web</td>
<td>2018</td>
<td><a href="https://bench.cs.vt.edu/deepargs">https://bench.cs.vt.edu/deepargs</a></td>
<td>Active</td>
</tr>
</tbody>
</table>

AR, antimicrobial resistance; ARG-ANNOT, Antibiotic Resistance Gene Annotation; CARD, Comprehensive Antibiotic Resistance Database; GROOT, Graphing Resistance Out of Metagenomes; HMM, hidden Markov model; NCBI, National Center for Biotechnology Information; PATRIC, Pathosystems Resource Integration Center; RGI, Resistance Gene Identifier; SEAR, Search Engine for Antimicrobial Resistance; ShortBRED, Short, Better Representative Extract Dataset; WGS, whole-genome sequencing.

Horizontal gene transfer (HGT): Passage of resistance genes from one bacterium to another when neither bacteria is the parent or daughter cell. This process usually occurs through transduction, conjugation or transformation.

Metagenomes
Collections of genes from all organisms of a given habitat or sample.

Now starting its second iteration⁴¹, seeks to benchmark these assemblers on highly complex and close to real data sets for users. However, currently, there is no single assembler that stands out as the best one that would accurately reconstruct known genomes and capture the majority of the taxonomic diversity in real data sets. Both biological factors (such as sample source and microbial community structure) and technical factors (such as library preparation method, sequencing depth and sequencing platform choice) affect the ability of an assembler to generate accurate and larger contigs. Thus, it is recommended to apply multiple assemblers on a subset of samples to determine the best fit for a given data set.

Following assembly, genomic or metagenomic contigs are annotated for resistance determinants by predicting protein-coding regions on contigs and then comparing them against antimicrobial resistance reference databases using similarity-based search tools (for example, BLAST⁴⁴, USEARCH⁴⁵ or DIAMOND⁴⁶). Although pairwise alignment between the query and antimicrobial resistance reference sequences is the most commonly applied approach for characterizing the resistome from contigs, an inherent bias of databases towards human-associated organisms is reflected in prediction outputs, so choosing the appropriate databases to compare assembled contigs with reference sequences is imperative⁴⁷.
**a Sample collection and sequencing**

Culture-dependent method (isolate genomics)

- DNA extraction
- Library preparation
- DNA sequencing
- Quality filtering

Culture-independent method (metagenomics)

- DNA extraction
- Library preparation
- DNA sequencing
- Quality filtering

**b Assembly-based versus read-based approach**

Assembly-based approach:
- Computationally expensive and time consuming, particularly in resistome profiling of large complex communities
- Identification of both known and novel resistance genes that share low similarity with reference database; however, requires high genome coverage
- Genomic context such as regulatory and mobile element sequences can be captured

Read-based approach:
- Fast and less computationally demanding, enabling resistome analysis of large data sets
- Identification of resistance genes is dependent on completeness of the reference database of organisms under analysis
- Nearby genes and genomic context are missing; may lead to false positives due to spurious mapping

**c Downstream analysis**

Ordination using resistance genes

- Antibiotic resistance gene families
- Samples

Resistome cluster analysis

- Antibiotic resistance gene families
- Samples

Antibiotic resistance gene network analysis

- Ceftazidime
- Amoxicillin and Clavulanic Acid
- Aztreonam
- Cefoxitin
- Tetracycline
- Chloramphenicol

Resistance profiling

- Resistance genes
- Samples

Fig. 2 | Assembly versus read mapping. a | The process for sequencing data generation for metagenomic and genomic samples. b | Steps for the read-based and the assembly-based methods of in silico resistance gene identification are contrasted. c | Examples of analysis that can be conducted on samples after resistance gene identification. CARD, Comprehensive Antibiotic Resistance Database.
Given sufficient coverage, assembly-based methods can construct whole genomes or large contigs with protein-coding genes, regulatory sequence information and the complete surrounding genomic context. This information can be used to study co-associated genes and biological pathways that are involved in resistance determination. Assembly and annotation of WMS data can identify antimicrobial resistance genes that are more divergent from and lack homology to known sequences in the reference databases. However, the process of de novo assembly and annotation is computationally expensive, time consuming and requires higher genome coverage than reference-based assembly or read mapping-based methods, which can be difficult to achieve for all samples, specifically when dealing with metagenomic samples with high microbial diversity and uneven taxonomic composition.

Read-based methods. Antimicrobial resistance genes in a sample can be detected without genome assembly either by aligning reads to the reference databases using pairwise alignment tools such as Bowtie2 (REF 1) or BWA49, or by splitting reads into k-mers and mapping them to the reference databases. SRST2 (REF 1) is one widely used tool that aligns reads to a custom reference database using Bowtie2 to predict antimicrobial resistance genes in the sample. Alternatively, KmerResistance1 splits reads into k-mers, maps them and counts the co-occurrence of k-mers between reads and a reference database to predict resistance genes and associated species. Both methods can identify antimicrobial resistance genes even in the presence of contaminants (for example, background noise in the raw reads owing to the presence of laboratory or host contamination) and in samples for which insufficient reads are available for de novo assembly, but they cannot predict antimicrobial resistance conferred by single-nucleotide polymorphisms (SNPs). By contrast, ARIBA39 uses a hybrid approach where reference sequences in the database are first clustered using CD-HIT53 before sequences from each cluster are assembled independently. Resulting contigs are then compared with the closest reference to identify allelic variants. Additionally, ARIBA provides information on whether genes are complete or fragmented and reports sequence variants along with their potential effects (for example, missense, nonsense or frameshift mutations and small insertions and deletions (indels)). Clustering reference sequences and using a representative sequence from the cluster to map reads considerably reduce ambiguous alignments48, but using a single linear representative locus masks subtle yet important variation between subtypes and subfamilies of genes within clusters34,45,49. To account for this information loss, Graphing Resistance Out of Metagenomes (GROOT)10, a newly established tool for resistome profiling of metagenomes, builds a variation graph for reference gene sets and aligns sequence reads to these graphs. Variation graphs are bidirectional acyclic sequence graphs that represent overall sequence variation within a given population. The alignment of reads against variation graphs effectively removes reference bias and facilitates accurate annotation of antimicrobial resistance genes. Before aligning sequences against a variation graph, traversals within the graphs are indexed by either Burrows–Wheeler transform or hash-map (minHash), indexing algorithms that considerably improve the mapping rate of large-scale sequencing reads to the graphs49,60. The read-based approach is generally fast and less computationally demanding because it bypasses de novo assembly, protein-coding gene prediction and pairwise alignment to public databases. For this reason, read-based methods have gained traction in recent years, especially in clinical diagnostics where conducting real-time sequencing-based resistance prediction is crucial.

Choosing the right approach. Presently, there is no consensus on which sequence analysis approach is better, and the choice of analysis mainly depends on the type of sequencing (WGS versus WMS), availability of computational resources and the study objective. Both approaches have trade-offs, as assembly causes information loss compared with direct read analysis48 but enables identification of protein-coding genes and for investigation of upstream and downstream regulatory elements, whereas direct read analysis lacks the positional information required to analyse upstream and downstream factors of identified resistance genes. New sequencing technologies, such as long-read sequencing and chromosome conformation capture-derived assays, are helping to alleviate this information loss by improving assembly fidelity (BOX 2).

The read-based approaches scale well with ever-increasing query sequences and antimicrobial resistance reference data. More importantly, they enable identification of antimicrobial resistance genes from low-abundance organisms present in complex communities, which may be missed by assembly-based methods owing to incomplete or poor assemblies. However, mapping reads directly to large data sets can inflate false-positive predictions, as reads derived from protein-coding sequences may spuriously align to other genes as a result of local sequence homology59. Thus, it is important that the reference databases are comprehensive and contain all variants of the reference genes. Database choice is especially critical when identifying antimicrobial resistance genes from large and complex communities such as soil and ocean, as novel or distant homologues of antimicrobial resistance genes present in understudied, less characterized environmental communities may be missed.

Well-studied sample types, such as the human gut, are now extensively characterized, even for low-abundance microorganisms and, thus, read-based approaches can be more confidently applied48. However, analysis of diverse samples is confounded by the lack of reference sequences, so the antimicrobial resistance genes in these environments are likely underestimated. To address this problem, a marker-based method, Short, Better Representative Extract Dataset (ShortBRED)91, was developed that enables fast and accurate profiling of the resistome in metagenomic data sets. ShortBRED first identifies marker sequences (short peptide sequences) representative of antimicrobial resistance protein families from the reference database and then maps reads
Box 2 | Sequencing-based innovations

Chromosome conformation capture assays
Chromosome conformation capture uses crosslinking, ligation and short-read sequencing to understand the spatial relationships of genetic material within cells\(^{28,30}\). These methods have been applied effectively in bacterial isolates and in metagenomic samples\(^{153,154}\). Using spatial information from chromosome conformation capture, researchers can increase their understanding of antimicrobial resistance gene regulation and begin to untangle more complicated resistance mechanisms. For example, these data may help resolve epistatic antimicrobial resistance gene relationships because gene spatial colocalization may indicate coordinated function. Chromosome conformation capture is also useful for identifying plasmid genes from chromosomal genes. This is important for antimicrobial resistance in particular because of the role of plasmids in horizontal gene transfer (HGT). Perhaps most interesting is the applicability of chromosome conformation capture to metagenomic samples. In addition to the points mentioned above, crosslinking information in metagenomic samples greatly reduces the complexity of metagenomic assembly because it enables researchers to group sequences by cell. Thus, antimicrobial resistance annotation methods that rely on assembled contigs are more viable if chromosome conformation capture information is also available.

Long-read sequencing
Long-read sequencing by single-molecule real-time sequencing (SMRT-seq) or nanopore sequencing can yield reads from 10 kb to over 100 kb (REFS\(^{16,17}\)). These long reads have several advantages over short reads for studying antimicrobial resistance in bacterial isolates and metagenomic samples. Long-read sequencing can greatly reduce the complexity of assembly for both bacterial isolates and metagenomic samples and, in some cases, can provide finished bacterial genomes (higher error rates than short reads means that sometimes a combination of short-read and long-read sequencing is needed to achieve these results\(^{29,31}\)). Better assemblies enable stronger interpretation of genomic context, providing similar benefits as with chromosome capture. Also similar to chromosome capture, long-read sequencing provides easier resolution of plasmid sequences, enabling more in-depth studies to understand antimicrobial resistance HGT. Additionally, both approaches give information about DNA methylation, which can inform genome assembly in metagenomic samples\(^{162,163}\).

Finally, long-read sequencing run times tend to be faster than short-read sequencing. This is useful for potential clinical deployments where speed may affect patient outcomes. Nanopore sequencing in particular excels in this area owing to its real-time data; these real-time data have already been used to provide rapid (within minutes) antimicrobial resistance gene-agnostic antimicrobial susceptibility testing (AST) predictions in Streptococcus pneumoniae\(^{164}\).

Transcriptomics
Transcriptomics techniques such as RNA sequencing (RNA-seq) facilitate analysis of bacterial gene expression, and these expression data have the potential to fill antimicrobial resistance knowledge gaps\(^{165,166}\). One major area where transcriptional data can contribute to antimicrobial resistance understanding is in connecting genotypic antimicrobial resistance data to phenotypic antimicrobial resistance results\(^{167}\). In cases where antimicrobial resistance genes are present but no resistance phenotype is found or in cases where two obvious antimicrobial resistance genes are present but antimicrobial resistance is confirmed, transcriptomic data may bridge the knowledge gap and point to novel antimicrobial resistance genes\(^{168,169}\).

Transcriptomics also has the potential to identify combinatorial gene effects resulting in antimicrobial resistance\(^{170}\). Finally, transcriptomics can successfully identify cases in which non-coding regulatory RNAs result in resistant bacterial phenotypes\(^{168,170}\).

Metagenomic assembly
Deconvolution and assembly of sequencing reads from a metagenomic sample.

Antimicrobial resistance databases
Both assembly-based and read-based approaches for the computational prediction of antimicrobial resistance in pathogens and environmental bacteria depend largely on curated antimicrobial resistance gene databases that link known genetic determinants of resistance to the antimicrobials they confer phenotypic resistance against (TABLE 2). These databases usually represent information accumulated from multiple studies that include AST of bacteria harbouring specific antimicrobial resistance genes.

Generalized versus specialized databases.
Public databases vary considerably in the scope of the resistance mechanisms\(^{171}\) that they cover and in the type of information they provide for annotations. Generalized antimicrobial resistance databases, such as the now archived Antibiotic Resistance Genes Database (ARDB)\(^{172}\) or the active Antibiotic Resistance Gene Annotation (ARG-ANNOT)\(^{173}\) and Comprehensive Antibiotic Resistance Database (CARD)\(^{174}\), cover broad spectrums of antimicrobial resistance genes and mechanism information, whereas specialized antimicrobial resistance databases provide comprehensive information for specific gene families or species (TABLE 2). For example, targeted databases such as Lactamase Engineering Database (LacED)\(^{175}\), the Lahey database of β-lactamases\(^{176}\), National Center for Biotechnology Information (NCBI) β-Lactamase Alleles Initiative, and the Comprehensive β-Lactamase Molecular Annotation Resource (CBMAR)\(^{177}\) focus on β-lactamases, a family of antimicrobial resistance enzymes that facilitate hydrolysis of the key β-lactam rings in β-lactam antimicrobials, thus protecting the bacteria from the antimicrobial activity. Resfinder\(^{178}\) is a web-based and standalone tool for detecting acquired antimicrobial resistance genes from sequenced or partially sequenced bacterial isolates (TABLE 1). Unlike other databases that require contigs as an input, Resfinder\(^{178}\) also accepts short reads as an input for comparison against known acquired resistance genes in bacterial genomes. In 2017, Resfinder\(^{178}\) updated its web-based service to enable identification of chromosomal mutations using PointFinder\(^{179}\). However, the identification of antimicrobial resistance-conferring chromosomal mutations is available for only a limited set of pathogenic microorganisms (Campylobacter, Escherichia coli, Mycobacterium tuberculosis, Neisseria gonorrhoeae, Plasmodium falciparum and Salmonella). Similar to Resfinder\(^{178}\), CARD\(^{180}\) offers its own tool, known as Resistance Gene Identifier (RGI), which uses curated antimicrobial resistance detection models to predict intrinsic antimicrobial resistance genes, dedicated resistance genes and acquired resistance from mutations in drug targets. RGI uses two antimicrobial resistance detection models: Protein Homologue Model for detecting functional homologues of antimicrobial resistance proteins and Protein Variant Model for the detection of mutations conferring antimicrobial resistance in otherwise sensitive targets. ARGs-OAP (v2)\(^{181}\) uses a custom database constructed from ARDB\(^{182}\) and CARD\(^{183}\), called SARG, with a hybrid UBLAST and BLASTYX algorithm, reflecting the critical need for a comprehensive
Table 2 | Summary of antimicrobial resistance reference databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
<th>Link</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General databases</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CARD²⁶</td>
<td>Ontology-based database that provides comprehensive information of AR genes and their resistance mechanisms</td>
<td><a href="https://card.mcmaster.ca/">https://card.mcmaster.ca/</a></td>
<td>Active; launched in 2013; updated monthly</td>
</tr>
<tr>
<td>SARG (v2)²⁷</td>
<td>Currently contains &gt;2,700 protein homologues and includes a curated set of resistance-conferring chromosomal mutations in protein-coding genes</td>
<td><a href="http://sarg.scripps.edu/">http://sarg.scripps.edu/</a></td>
<td>Active; started in 2012; update regularly; last update in February 2019</td>
</tr>
<tr>
<td>Resfinder²⁸</td>
<td>Collection of AR genes involved in HGT events</td>
<td><a href="https://cge.cbs.dtu.dk/services/ResFinder/">https://cge.cbs.dtu.dk/services/ResFinder/</a></td>
<td>Active; started in November 2016</td>
</tr>
<tr>
<td>ARG-ANNOT²⁹</td>
<td>First centralized resource of AR gene information</td>
<td><a href="https://arg-annot.vital-it.org/">https://arg-annot.vital-it.org/</a></td>
<td>Archived; last update in January 2015</td>
</tr>
<tr>
<td>ARDB²⁵</td>
<td>Collation of multiple databases (CARD, ARG-ANNOT and ResFinder) to avoid redundancy between entries</td>
<td><a href="https://ardb.cbcb.umd.edu/">https://ardb.cbcb.umd.edu/</a></td>
<td>Active; last update in December 2016</td>
</tr>
<tr>
<td>MEGARes²⁷</td>
<td>Collation of resistance gene variants identified in multiple functional metagenomics studies</td>
<td><a href="http://megares.meglab.org/">http://megares.meglab.org/</a></td>
<td>Active; started in 2016</td>
</tr>
<tr>
<td>ARG-ANNOT²⁹</td>
<td>Repository of &gt;1,800 AR sequences collated from scientific literature and online resources</td>
<td>Not available</td>
<td>Archived; last update in May 2018</td>
</tr>
<tr>
<td>Mustard⁸⁵</td>
<td>Resource containing 6,095 AR determinants from 20 families, including curated sets of AR genes identified in functional metagenomics studies</td>
<td><a href="http://mgps.eu/Mustard/">http://mgps.eu/Mustard/</a></td>
<td>Active; last update in November 2018</td>
</tr>
<tr>
<td>FARME database⁵⁴</td>
<td>Collated set of microbial sequences functionally screened to confer resistance in various functional metagenomics studies of different habitats</td>
<td><a href="http://staff.washington.edu/jwallace/farme/">http://staff.washington.edu/jwallace/farme/</a></td>
<td>Active; last update in 2017</td>
</tr>
<tr>
<td>SARG (v2)²⁹</td>
<td>Hierarchical structured database derived from ARDB, CARD and NCBI-NR database</td>
<td><a href="http://smile.hku.hk/SARGS">http://smile.hku.hk/SARGS</a></td>
<td>Active</td>
</tr>
<tr>
<td>Lahey list of β-lactamases⁷⁶</td>
<td>First initiative to compile known β-lactamases and assign nomenclature to novel ones</td>
<td><a href="http://www.lahey.org/Studies/">http://www.lahey.org/Studies/</a></td>
<td>Archived; last update in 2015</td>
</tr>
<tr>
<td>BLDB¹⁷⁸</td>
<td>Manually curated database for AR enzymes classified by class, family and subfamily</td>
<td><a href="http://bldb.eu/">http://bldb.eu/</a></td>
<td>Active; last update in November 2018</td>
</tr>
<tr>
<td>LacED⁹⁶⁶</td>
<td>Curated database of TEM and SHV β-lactamases, including a curated set of known TEM and SHV variants</td>
<td><a href="http://www.laced.uni-stuttgart.de/">http://www.laced.uni-stuttgart.de/</a></td>
<td>TEM LacED active: last update in 2017; SHVED archived: last update in April 2010</td>
</tr>
<tr>
<td>CBMAR⁷¹</td>
<td>Database that identifies and characterizes novel β-lactamases on the basis of Ambler classification</td>
<td><a href="http://proteininformatics.org/mkumar/lactamasedb/">http://proteininformatics.org/mkumar/lactamasedb/</a></td>
<td>Last update in September 2014</td>
</tr>
<tr>
<td><strong>Species-specific databases</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUBII-TB-DB²⁶</td>
<td>Database of mutations associated with AR in Mycobacterium tuberculosis</td>
<td><a href="https://umr5558-bibiserv.univ-lyon1.fr/mubii/mubii-select.cgi">https://umr5558-bibiserv.univ-lyon1.fr/mubii/mubii-select.cgi</a></td>
<td>Last update in December 2013</td>
</tr>
</tbody>
</table>


Species-specific databases exist for pathogenic or model bacteria such as *M. tuberculosis* (for example, Tuberculosis Drug Resistance Database³¹ or MUBII-TB-DB²⁶) and *E. coli* (TABLE 2). These species-specific databases are invaluable for understanding resistance in these specific organisms but also highlight the importance of considering antimicrobial resistance genes in...
their phylogenetic context, especially as some bacteria can have intrinsic resistance to some antimicrobials (reviewed previously\(^7\)). Species-centric databases enable rapid and effective curation of new antimicrobial resistance genes and chromosomal mutations and can offer quick preliminary screening for characterization. Such screening has proved highly effective for pathogens such as *M. tuberculosis* in which HGT events are rare and drug resistance originates mainly from chromosomal mutations\(^8,9\). The CRyPTIC Consortium and 100,000 Genomes Project demonstrated this effectiveness in *M. tuberculosis* with resistance predictions with over 90% sensitivity and specificity for all four first-line anti-tuberculosis drugs\(^10\).

While these tools are all steps in the right direction, a continuously updating and comprehensive database with extensive gene metadata and the ability to find both point mutation matches and remote homologues is needed.

### Hidden Markov model-based databases

One major limitation of these databases is that the antimicrobial resistance genes they contain are heavily biased towards human pathogens and easily cultivable model organisms, making it difficult to identify remote homologues or novel resistance sequences present in fastidious or uncultured bacteria\(^11\). This bias complicates antimicrobial resistance gene identification across less commonly studied bacteria, a difficulty that is magnified by the diverse and complicated mechanisms that cause resistance\(^12\). One potential solution to overcome this bias is to use hidden Markov model (HMM) databases. Derived from the multiple sequence alignment of known sequences, an HMM can find sequences with similar function but low sequence identity\(^14,15\). Resfams\(^16\) is an HMM database of antimicrobial resistance proteins derived from multiple sequence alignments of manually curated sets of representative antimicrobial resistance protein sequences obtained from the generalized CARD and the specialized LaED\(^17,18\) and Lahey database\(^19\) (TABLE 2). The authors of the Resfams\(^16\) database showed that it can identify a substantially greater number of novel antimicrobial resistance genes and remote homologues of known antimicrobial resistance genes than other databases such as ARDB and CARD that rely on BLAST-based methods for gene identification. A direct comparison of manually curated antimicrobial resistance gene sets showed that Resfams\(^16\) identified 64% more antimicrobial resistance genes in both soil and human gut microbiota than the BLAST-based search of CARD and ARDB. This increased sensitivity demonstrates the versatility of the HMM in annotating sequences from non-clinical samples with sparser representation in publicly available resistance gene databases. However, HMM-based approaches may have poor specificity (yield higher number of false-positive hits) and may not be able to distinguish between protein families with closely related functions. This could occur owing to the higher probability of selecting sequences from other subfamilies on the basis of domains common to the family. To mitigate the lack of specificity, Resfams\(^16\) (like the Pfam database) uses curated thresholds (for example, a gathering threshold) for each profile HMM. These profile-specific gathering threshold values set an inclusion or exclusion bit score cut-off by comparing it with test data sets containing negative sequences. Currently, Resfams\(^16\) contains 166 profile HMMs that represent major antimicrobial resistance gene families. HMM-based antimicrobial resistance databases could be valuable in identifying large and diverse arrays of resistance determinants in understudied environmental samples compared with BLAST-based databases. However, current HMM-based databases do not identify resistance arising from chromosomal mutations. To further facilitate the detection of antimicrobial resistance genes in large complex environments, the Functional Antibiotic Resistance Metagenomic Element (FARME)\(^20\) database comprises a curated set of microbial sequences excluded from current databases but functionally screened to confer resistance in various functional metagenomics studies of different habitats. Apart from predicted protein-coding antimicrobial resistance sequences, the FARME database also includes regulatory elements, mobile genetic elements and predicted proteins flanking antimicrobial resistance genes.

A similar database, the functional resistance database (ResfinderFG)\(^21\), was built by aggregating data from four functional metagenomics studies selected against 23 antimicrobials. When comparing this database with the Resfinder\(^-\) database, the authors noted that they found different results by total antimicrobial use; this observation may represent a difference in how resistance is conferred when putative resistance determinants are cloned into *E. coli* as compared with when they are expressed in their native bacterial host.

The Mustard\(^22\) antimicrobial resistance determinants database uses an innovative approach of incorporating 3D protein structure to help predict resistance genes. When this approach was applied to predicted proteins from metagenomic samples, it predicted >6,000 resistance genes compared with 67 genes identified by BLASTP and 50 by Resfinder\(^-\), suggesting higher sensitivity.

### Remaining challenges

Considerable developments in biocuration of antimicrobial resistance sequences have enabled the identification and characterization of antimicrobial resistance genes from genomes and metagenomes, but several limitations still preclude cost-effective and rapid antimicrobial resistance surveillance. One major bottleneck is the lack of effective curation strategies. With few exceptions, antimicrobial resistance databases lack efficient and sustainable curation pipelines, so they tend to receive active maintenance for a few years before becoming outdated.

Many antimicrobial resistance genes can be assigned names on the basis of nucleotide sequences and protein sequences, leading to conflicting naming schemes. Conflicting gene names and synonyms create redundancy across databases and confuse users (for example, dihydrofolate reductase is referred to as *dhfr* in some databases and *dfrA* in others)\(^26\). This problem is
Antimicrobial resistance genomic data are an ever-expanding data source. HGT events and selection pressures that proliferate new antimicrobial resistance mutations require active biocuration strategies whereby entries can be curated as they are recognized. The propagation of new colistin resistance mechanisms such as \textit{mcr-1}, which was first described in 2016 from Chinese bacterial isolates\cite{1} and then subsequently identified worldwide in newly collected and previously stored isolates\cite{2,3}, demonstrates the need for frequent database updating and curation. When properly implemented, they facilitate rapid collection of epidemiological data for recently discovered resistant determinants\cite{4}. Indeed, antimicrobial resistance annotation should be a continuous effort, as all downstream analyses depend on the accuracy of reference databases. Establishing best practices for biocuration, systematically assigning annotations to newly discovered genes and preventing misinterpretations will pay dividends for public health and basic science.

Another important limitation of current antimicrobial resistance databases is their focus on the identification and characterization of protein-coding resistance genes; they ignore other potential antimicrobial resistance mechanisms such as genomic changes or de novo mutations in ribosomal RNA (rRNA) genes and regulatory elements and drug target mutations. Recent efforts by CARD\cite{5} and Resfinder\cite{6} have tried to address this issue.

**Functional metagenomics**

In addition to sequence-based metagenomics, functional metagenomics is a powerful, culture-independent, sequence-unbiased approach for characterizing resistomes\cite{7,8}. In this method, a metagenomic library is generated by cloning the total community DNA extracted from a sample into an expression vector. This library is transformed into a susceptible indicator host strain and is assayed for antimicrobial resistance by plating on selective media that are lethal to the wild-type host. The selected inserts from the surviving recombinant, antimicrobial-resistant host cells are then sequenced, and resulting sequences are subsequently assembled and annotated (Fig. 3). Parallel Annotation and Reassembly of Functional Metagenomic Selections (PARFuMS)\cite{9} is a custom computational pipeline that assembles reads from functional metagenomic selections into contigs using the Velvet\cite{10} and Phrap\cite{11} assemblers and annotates the assemblies for antimicrobial resistance genes using MetaGeneMark\cite{12} and Resfams\cite{13}. This approach enables high-throughput analysis of large genomic content (up to 50 Gbp of unique metagenomic DNA interrogated per library), and antimicrobial resistance phenotypes can be associated directly with causative genes, obviating the need to culture individual antimicrobial resistance gene carriers.

Functional metagenomics has enabled the discovery of several new antimicrobial resistance mechanisms and their related genes\cite{14}. One such example is the recently discovered tetracycline resistance mechanism by tetracycline destructases\cite{15}, whereby soil functional metagenomics led to the discovery of nine genes that confer tetracycline resistance through enzymatic inactivation. Further analysis and biochemical characterization revealed that these enzymes catalyse tetracycline oxidation in an FAD-dependent manner, thereby inactivating tetracycline\cite{16}.

While the preceding study shows the strength and usefulness of functional metagenomics, this approach has certain limitations. For example, a gene has to be functional outside its native microbial host to be identified by functional metagenomic selections. Many times, differences between a recombinant expression host such as \textit{E. coli} and the original host (for example, some Gram-positive organisms) do not confer the same phenotype for the same gene. This problem was highlighted by studies showing effects of different hosts on the same metagenomic libraries\cite{17,18}. Thus, there is a need to include a phylogenetically diverse group of hosts that can be used for functional metagenomic selections. In addition, genes outside their genomic context, such as syntenic regulatory elements, may have different phenotypes in the recombinant expression host from those in the original host\cite{19}. Thus, it is important that novel antimicrobial resistance genes identified by functional metagenomics screens be characterized microbiologically and biochemically. Extension of the current functional metagenomics approach and development of new techniques to discover novel resistance genes are deserving research directions.

**Machine learning for resistance prediction**

Numerous studies have explored machine learning algorithms for studying antimicrobial resistance, highlighting its role in predicting resistance phenotype directly from genotype. Machine learning approaches can be implemented as supervised learning or unsupervised learning approaches. In supervised learning, the training data set with outcome of interest can be utilized to build a prediction model that can be further applied to query sequences to predict their outcome. Several studies have used gene presence or absence and AST outcomes as features to create the training set for models. In one study, a logistic regression approach was used to develop a model based on 14 gene parameters and 3 molecular typing markers that can differentiate between vancomycin-susceptible and vancomycin-resistant \textit{Staphylococcus aureus} using publicly available genomic data and patient isolates\cite{20}. The model performance was tested by a leave-one-out validation method, and it showed 84% classification accuracy. Although this accuracy level does not meet clinical standards, the approach provides an important proof of concept that motivates the development of more sophisticated models for identifying antimicrobial resistance. Another study evaluated a rules-based and a machine learning-based approach (that is, logistic regression) for predicting antimicrobial resistance profiles and showed that the machine learning-based approach had higher accuracy with novel variants.
Fig. 3 | Functional metagenomics to interrogate acquired resistance genes in different environments and human pathogens. A summary of experimental and computational steps involved in functional metagenomics. First, sample collection and extraction occur. Metagenomic DNA is isolated from the sample (for example, soil or faeces). Second, functional selection using an expression vector and the host system is performed. The metagenomic DNA is sheared to a target size of 2–5 kb, and the fragments are then cloned into an expression vector and transformed into a host system (for example, Escherichia coli). The transformants are then selected using antimicrobials at concentrations that are inhibitory to the wild-type host system. Third, barcoded sequencing of pooled DNA fragments is performed. The resistance-conferring fragments are PCR amplified, barcoded and pooled together for sequencing. The sequencing reads are computationally demultiplexed using barcode assembly and quality trimmed to obtain high-quality clean reads. Fourth, iterative assembly of sequencing reads by sample is performed. The clean reads are assembled with computational pipeline Parallel Annotation and Reassembly of Functional Metagenomic Selections (PARFuMS), in which ensemble-based assembly is performed using multiple rounds of a short-read assembler (Velvet), and intermediate contigs are then used in a long-read assembler (Phrap) to give full-length contigs. Finally, resistance gene annotation of assembled reads is performed. The annotation of contigs is accomplished using BLAST-based and hidden Markov model (HMM)-based databases.
Carbapenem resistance

Resistance against the broad-spectrum carbapenem class of β-lactam antimicrobials, which are often used as drugs of last resort.

Methicillin resistance

Resistance against methicillin, a narrow-spectrum penicillin derivative. Methicillin resistance is often seen in the context of methicillin-resistant Staphylococcus aureus (MRSA), a common human pathogen. This resistance is commonly gained by horizontal transfer of a modified target protein (see Fig. 1b and 1c).

Deep learning

An extension of representational machine learning methods where the algorithm uses multiple transformation layers between raw data and output rather than one layer. This often improves results for more complex machine learning tasks.

in known antimicrobial resistance genes than the rules-based approach.

Recent studies and tools use k-mers derived from whole genomes of antimicrobial-resistant and antimicrobial-susceptible species along with their AST outcomes to develop prediction models. Mykrobe predictor, a fast k-mer screening tool, is used to identify antimicrobial resistance genes and SNPs in S. aureus and M. tuberculosis. It utilizes the curated genetic information of resistant and susceptible alleles of the same species to build reference graphs (DBG) of these two categories and to map k-mers derived from sequencing reads to these graphs. Mykrobe predictor showed 99.1% and 82.6% sensitivity and 99.6% and 98.5% specificity for S. aureus and M. tuberculosis, respectively, on an independent validation set and provided important insights on potential antimicrobial resistance elements.

By contrast, Rapid Annotation using Subsystem Technology (RAST) is a k-mer-based tool that uses a machine learning classifier (AdaBoost) based on the Pathosystems Resource Integration Center (PATRIC) database to identify target-specific antimicrobial resistance genes in a specific collection of pathogens. RAST is trained on k-mer data derived from the contigs of each genome. These k-mer counts were converted into a binary matrix of 1s and 0s to depict whether a particular k-mer is present in that genome or not. The binary matrix along with AST outcome is then used to form a classifier model as well as to identify putative k-mers associated with resistance. The RAST classifier could identify carbapenem resistance in Acinetobacter baumannii, methicillin resistance in S. aureus and β-lactam and co-trimoxazole resistance in Streptococcus pneumoniae with accuracies of 88–99%.

One major shortcoming of any machine learning classifier is its dependency on the training data or existing knowledge base. To apply machine learning classifiers in clinical diagnostics, a large data set of curated antimicrobial resistance genes that contains accurate genotypic data linked to curated AST data will be required to build an effective and robust machine learning-based classifier for antimicrobial-resistant organisms. In addition to differentiating between an antimicrobial-resistant and antimicrobial-susceptible organism, machine learning approaches are currently being applied to predict antimicrobial resistance genes in metagenomic data. DeepArgSS10 is a newly established tool that applies deep learning to identify antimicrobial resistance genes. On the basis of curated data sets of CARD and ARDB combined with Uniprot protein data, DeepArgSS built a dissimilarity matrix between antimicrobial resistance proteins and non-antimicrobial resistance proteins and used it to train two deep-learning models: DeepArg-LS for assembled genes and DeepArg-SS for short reads. These models can be used to predict antimicrobial resistance genes in new test data.

Although the application of machine learning to antimicrobial resistance prediction and classification is promising, these techniques have a long way to go before they can be used for rapid diagnostic purposes and replace traditional culture techniques and AST, which can take days or weeks to yield results.

Conclusions and future perspectives

Antimicrobial resistance is a major public health threat. Monitoring and understanding the prevalence, mechanisms and spread of antimicrobial resistance are priorities for both individual patient care and global infection control strategies. Despite stellar advancements, hurdles for antimicrobial resistance detection and understanding persist. Costs are decreasing for sequencing and for automated antimicrobial resistance detection instruments, but start-up and operation costs still outstrip many health-care budgets. Further cost reductions for these technologies will be important for widespread adoption.

The accurate identification of resistance determinants and the correlation of antimicrobial resistance gene profiles to antimicrobial treatment outcomes will facilitate personalized approaches to developing treatment regimens. The success of this approach depends heavily on the comprehensiveness and quality of public antimicrobial resistance gene databases, which have major roles in the development of biological assays and computational tools that expand our ability to detect resistance genes in single isolates and in microbial communities. While progress has been made in building comprehensive antimicrobial resistance gene databases, lack of standardization across databases and long update intervals hold back their potential. Moreover, complex resistance mechanisms (Fig. 1b) are difficult to capture in antimicrobial resistance databases. For example, resistance can arise from epistatic relationships between multiple genes such as in the case of carbapenem resistance, which can arise from the combination of extended-spectrum β-lactamases and efflux pumps or porin impermeability. Resistance can even occur via overexpression of normal genes such as those encoding efflux pumps, and detection of these resistance mechanisms requires transcriptional measurements ([Fig. 1b,c]) on the basis of curated data sets of CARD and ARDB combined with Uniprot protein data, DeepArgSS built a dissimilarity matrix between antimicrobial resistance proteins and non-antimicrobial resistance proteins and used it to train two deep-learning models: DeepArg-LS for assembled genes and DeepArg-SS for short reads. These models can be used to predict antimicrobial resistance genes in new test data.

Although techniques for novel antimicrobial resistance gene discovery exist, such as functional metagenomics, these techniques still have major caveats in
the types of antimicrobial resistance genes that they can detect. Innovative methods to determine other antimicrobial resistance gene mechanisms are sorely needed. Moreover, robust models to predict which resistance genes will spread both on the local level within a health-care setting and on the global scale between countries are needed. These models will likely need to incorporate not only the antimicrobial resistance gene sequence and mechanism but also the genomic context, host bacterial species and geographic location.

Rapid and accurate identification of resistance genes in isolate and metagenomic samples would augment the ability of clinicians to make treatment plans for bacterial infections, facilitating a future where sequence-based personalized medicine is routine. It would also ease antimicrobial resistance surveillance efforts and enable low-resource areas to benefit more fully from rapidly decreasing sequencing costs.

Published online: 18 March 2019