Journal of Microbiology (2021) Vol. 59, No. 3, pp. 270–280
DOI 10.1007/s12275-021-0652-4
eISSN 1976-3794
pISSN 1225-8873

REVIEW

Overview of bioinformatic methods for analysis of antibiotic resistome from genome and metagenome data

Kihyun Lee[1,2], Dae-Wi Kim[3], and Chang-Jun Cha[1*]

[1]*Department of Systems Biotechnology and Center for Antibiotic Resistome, Chung-Ang University, Anseong 17546, Republic of Korea*
[2]*ChunLab, Inc., Seoul 06194, Republic of Korea*
[3]*Division of Life Sciences, Jeonbuk National University, Jeonju 54896, Republic of Korea*

(Received Dec 11, 2020 / Revised Jan 28, 2021 / Accepted Jan 29, 2021)

Whole genome and metagenome sequencing are powerful approaches that enable comprehensive cataloging and profiling of antibiotic resistance genes at scales ranging from a single clinical isolate to ecosystems. Recent studies deal with genomic and metagenomic data sets at larger scales; therefore, designing computational workflows that provide high efficiency and accuracy is becoming more important. In this review, we summarize the computational workflows used in the research field of antibiotic resistome based on genome or metagenome sequencing. We introduce workflows, software tools, and data resources that have been successfully employed in this rapidly developing field. The workflow described in this review can be used to list the known antibiotic resistance genes from genomes and metagenomes, quantitatively profile them, and investigate the epidemiological and evolutionary contexts behind their emergence and transmission. We also discuss how novel antibiotic resistance genes can be discovered and how the association between the resistome and mobilome can be explored.

Keywords: antimicrobial resistance, antibiotic resistome, genome, metagenome

Introduction

Methods for the generation and analysis of whole-genome sequencing (WGS) data have been actively developed over the last few decades. A recently proposed experimental design for the large-scale WGS-based surveillance of bacterial pathogens enables a sequencing cost of 10 USD per strain (Perez-Sepulveda *et al.*, 2020). WGS has become an increasingly feasible choice in the laboratories studying clinical or antibiotic-resistant isolates, and studies based on WGS data have begun to unravel the global epidemiology and evolutionary processes behind the recent emergence of resistance against clinically important antibiotics in several pathogens.

Metagenomics approaches based on high-throughput sequencing have become popular tools for studying the distribution and dynamics of antibiotic resistance at the whole microbiome scale. A key advantage of the metagenomics approach is the capacity for unbiased cataloging of antibiotic resistance genes (ARGs), whereas there are some differences in the results obtained according to specific purposes ranging from analysis of known ARGs to discovery of completely novel ARGs. The results from these approaches can also include the landscape and ecological dynamics of ARGs in host-associated microbiomes, monitoring ARGs in sewage for regional surveillance and prediction, investigating human impact on the environmental distribution of ARGs at a local scale, elucidating the factors governing the prevalence of resistance in the global ecological context, and finding previously unreported ARGs.

In this review, we provide an overview of the currently popular strategies for the utilization of whole genome and metagenome sequencing data in the studies of antibiotic resistome. For the WGS-based approaches, we include all major steps from assembly to species identification, ARG detection, and epidemiological investigation. For the whole metagenome sequencing (WMS), we cover the approaches for cataloging and quantitative profiling of ARGs from microbiome samples. Identification of novel ARGs and comprehensive exploration of mobile ARGs are often considered as high-priority objectives in resistome studies. Hence, we also describe the available high-throughput sequencing-based approaches for the discovery of novel ARGs and the exploration of mobile resistome.

Genetic determinants of antibiotic resistance

Antibiotic resistance of a bacterium can be defined from two different perspectives, either from a clinical point of view based on treatment outcomes, or from a microbiological perspective based on the bacterial growth response under various antibiotic concentrations. In both cases, the acquisition of resistance to previously effective antibiotics is achieved mostly through mutation or horizontal gene transfer (HGT) occurring in its genome.

Various molecular and cellular mechanisms neutralize the

*For correspondence. E-mail: cjcha@cau.ac.kr
Copyright ⓒ 2021, The Microbiological Society of Korea

effect of antibiotics and confer antibiotic resistance phenotypes. Resistance can be derived mainly via (a) efflux pump, (b) modification of the antibiotic target protein by mutation, (c) enzymatic degradation (inactivation) of antibiotics, (d) overexpression of the proteins targeted by antibiotics, (e) metabolic bypasses by acquisition of alternative pathways, and (f) reduction in the cell permeability (Ellington *et al.*, 2017). Intrinsic resistance to certain antibiotics may exist in some bacterial clades when the antibiotic target may not exist, or the molecular or cellular pathways that neutralize the antibiotic action mechanism may be present in most members of a given clade.

Acquired resistance, which is manifested by a subset of members within the clade, is achieved by the genomic variations that result in the expression of the above-mentioned resistance mechanisms. The causal genomic variations of acquired resistance (i.e., the genetic determinant of antibiotic resistance) occur in a variety of forms, including point mutation and homologous recombination in protein-coding or noncoding sequences, structural variations such as gene amplification or rearrangement, and acquisition of the mobile genetic elements encoding the enzymes that convey one of the resistance mechanisms when expressed.

WGS-based analysis for antibiotic resistance

Cataloging and quantitative profiling of known antibiotic resistance determinants in a set of genomic or metagenomic sequences is conceptually straightforward, because it simply needs the comparison between known resistance determinants in the databases and the sequences from given samples. Figure 1 presents the schematic context for antibiotic resistome research using genomics and metagenomics approaches. Typical workflows in the WGS-based analysis of antibiotic resistance start with the assembly step, followed by species identification, strain subtyping, annotation of genes and mutations associated with antibiotic resistance, and the post-analyses for evolutionary processes. The following subsections address each of these steps. We will focus mostly on the known ARGs; therefore, the discovery of novel antibiotic determinants will not be addressed in detail.

Assembly

A common starting point in WGS data analysis workflow is read quality preprocessing and assembly. Assembly quality has an overarching influence on the downstream analyses, especially regarding the recovery of mobile genetic element

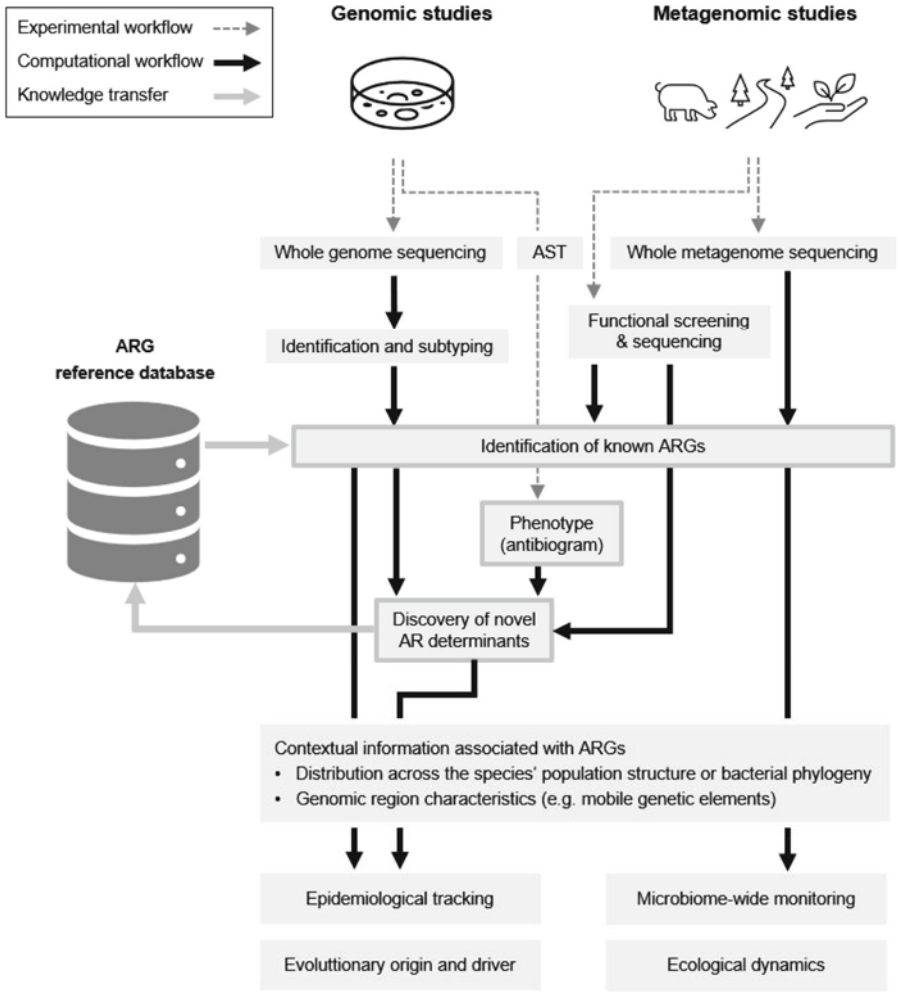

Fig. 1. Genomics and metagenomics approaches in antibiotic resistome research.

(MGE) structures that are frequently encountered around the horizontally acquired ARGs because the repetitive nature inherent to those MGEs hinders the assembly process. The achievement of high-quality assembly (i.e., contig generation with only a few remaining gaps, low error rate, and circularized plasmids) depends on multiple factors, including the genomic organization, sequencing strategy, and assembly method. According to the surveillance across NCBI RefSeq records, the most popular sequencing method in bacterial WGS has been Illumina platforms, followed by the PacBio platforms, whereas the WGS data generated from Nanopore platforms have been increasing recently (Segerman, 2020). The principal trade-off in the sequencing platform is between the scale of the isolate panel (i.e., the number of strains to sequence) versus the completeness of each assembly. Assembly workflows involve read pre-processing, assembly generation, and optionally the polishing of final assembly, and the choices of tools and parameters for the optimal performance depend on the sequencing platforms. Popular assembly tools used for bacterial genomes include SPAdes (Bankevich *et al.*, 2012) for short reads, and Flye (Kolmogorov *et al.*, 2019) and miniasm/minipolish (Wick and Holt, 2020) for long reads. For the extensive guides covering the pre-processing and polishing steps, we recommend a useful web page (github.com/rrwick/Trycycler/wiki/Guide-to-bacterial-genome-assembly), which provides an interactive guide for the optimal assembly procedure based on sequencing methods and user-specific priority.

Assembly-free workflows

There are certain cases where the assembly step is omitted to streamline the computational process. It is typically the case in studying genetically monomorphic pathogens with pangenomes such as *Mycobacterium tuberculosis,* where the strains are not assumed to display significant structural variations and the screening of point mutations is considered to be sufficient for cataloging the resistance determinants (Cohen *et al.*, 2019).

A series of analytical workflows is developed for analyzing raw reads generated from the metagenomes of clinical specimens (e.g., infected body fluids) using Nanopore technology, for which the key objectives are to achieve taxonomic identification, subtyping, and antibiotic resistance prediction at the genome level within a few hours of sequencing run (Břinda *et al.*, 2020). Due to the short turnaround time, nanopore sequencing-based analysis of clinical metagenomes is a promising point-of-care testing method that may support or replace cultivation-based diagnostic assays.

Species identification

Species-level identification sets the basis for subsequent steps, such as strain subtyping, choice of guideline for antibiotic susceptibility testing, identification of antibiotic resistance-associated mutations, and epidemiological and evolutionary interpretation. Workflow for species-level identification based on WGS data could be accelerated by narrowing down the candidate taxonomy of the query genome using rough but efficient identification methods. Typical options for such purposes include 16S rRNA gene sequence extraction and searching against reference 16S rRNA gene-based taxonomy databases such as EzBioCloud (Yoon *et al.*, 2017), extraction of universally conserved core genes and the subsequent search against reference databases (Chaumeil *et al.*, 2019), and the use of rapid whole-genome similarity calculators such as Mash (Ondov *et al.*, 2016) against the entire bacterial genomes. Conclusive identification can be drawn based on the pairwise average nucleotide identity values against the panel of reference genomes. Selection of reference genomes can be guided by data repositories specialized for prokaryotic type strains such as gcType (Shi *et al.*, 2020) and EzBioCloud (Yoon *et al.*, 2017). The entire workflow outlined above can also be accomplished using web-based tools such as TYGS (Meier-Kolthoff and Göker, 2019) and gcType (Shi *et al.*, 2020). For high-throughput applications with a large number of genomes, it might be more appropriate to use GTDB-Tk (Chaumeil *et al.*, 2019), which is locally executable, or in-house pipelines built by the user themselves.

Subtype identification at the strain level

Subtyping of strains using the schemes that are conventionally used in the field would provide highly informative epidemiological contexts. Importantly, epidemiological studies on resistant bacteria are often carried out in the subtype context. The genetic relationship at the subtype level has a strong association with variations in antibiotic resistance phenotypes (MacFadden *et al.*, 2020). Well-established subtyping schemes usually exist for extensively characterized pathogenic species, but not for most non-pathogenic species. Of the phenotype-based subtyping schemes, serotyping is probably the most widely used method across different pathogenic species. The WGS-based serotype prediction has been developed only for a few pathogens, including *Escherichia coli* (Joensen *et al.*, 2015), *Salmonella* spp. (Yoshida *et al.*, 2016), and *Klebsiella pneumoniae* (Wick *et al.*, 2018), but not for most pathogens. For the genotype-based typing schemes, multilocus sequence typing (MLST) is the most popular approach for pathogens. PubMLST functions as the central repository for MLST schemes and currently hosts the scheme databases for 127 taxa defined at species or higher ranks (Jolley *et al.*, 2018). Tools available for WGS-based MLST typing using the PubMLST schemes, such as MLSTcheck (Page *et al.*, 2016), allow high-throughput subtyping of WGS data.

Strain subtyping methods that utilize variations at the whole genome scale have been developed more recently. Whole-genome subtyping methods can be classified into core genome MLST (cgMLST), SNP-based methods, and genome-wide k-mer methods. cgMLST methods can be viewed simply as the extension of classical MLST, which uses many loci. Databases organized for the species-specific cgMLST schemes and isolate information are provided at EnteroBase (Zhou *et al.*, 2019) and the cgMLST.org server (Ridom GmbH). Recently, the approach developed for strain subtyping using k-mers, the PopPUNK method, uses variable-length k-mers to cluster the genomes based on sharing of core and accessory genome sequences (Lees *et al.*, 2019). PopPUNK is fast enough to be scalable over many input genomes. PopPUNK cluster numbers have already been utilized in the literature to refer to clones defined at high resolution. The effectiveness of a subtyping scheme as a means of providing epidemiological

contexts depends on the popularity of the scheme. The more studies use the same method and nomenclature, the more epidemiological information accumulates under the consistent rule. The whole-genome scale subtyping methods have not yet gained popularity as classical MLST and serotyping. However, considering that the field of WGS-based epidemiology still has a short history, it is expected that it will become increasingly popular and standardized in the upcoming years.

Identification of antibiotic resistance genes and mutations

Several databases are available that provide the collections of known ARGs. Databases, which are highly popular, actively maintained, and contain original curations, include the CARD (Alcock *et al.*, 2019), ResFinder (Bortolaia *et al.*, 2020), ARG-ANNOT (Gupta *et al.*, 2014), DeepARG (Arango-Argoty *et al.*, 2018), and AMRfinder (Feldgarden *et al.*, 2019) (the reference sequences of AMRfinder are also available at BioProject PRJNA313047). Most of these databases have an accompanying annotation tool designed specifically for the database. Annotation tools typically perform the sequnce alignment search and filtering at reliable cutoffs. For example, the Resistance Gene Identifier (RGI) annotation tool coupled with CARD uses blast bit score cutoffs manually curated to maximize the differentiation of each CARD entry against the background NCBI nr proteins (Alcock *et al.*, 2019) and these cutoff values can be referred to from the 'blastp_bit_score' field associated with each protein homolog model. The AMRfinder tool uses the bit score cutoffs manually curated for each of the profile hidden Markov models (HMM) constructed from the reference ARG families (Feldgarden *et al.*, 2019) and cutoff value information can be obtained from the TC (trusted cutoff) field in each HMM file of reference ARGs. DeepARG evaluates each blast alignment (i.e., a metagenomic read aligned to each reference ARG) by calculating the probability that it represents a true-positive hit using pre-trained deep learning models (Arango-Argoty *et al.*, 2018). Although DeepARG does not require bit score thresholds to be defined for each ARG, this tool allows users to set a probability threshold (e.g., 0.8) to filter out non-specific hits. Cutoffs for alignment identity, score, and coverage breadth should not be set with somewhat arbitrary decisions because these cutoffs often have a dramatic impact on the final list of ARGs detected. As such, it is highly desirable to use the software package provided in conjunction with the corresponding database to utilize the gene family-specific adjusted alignment cutoffs. Apart from the databases made up with the original curation, there are packages such as ABRICATE (https://github.com/tseemann/abricate) that integrated several independent databases and provided its own wrapper script for convenient one-stop annotation.

For certain species, it may be desirable to identify the point mutations that are known to cause resistance against certain antibiotics. A majority of known resistance mutations are located on the genes encoding central cellular functions such as *gyrA*, *rpoB*, or ribosomal RNAs. There are a number of species-specific curated databases available for known point mutations associated with antibiotic resistance. These databases also provide annotation tools for detecting target mutations in the input genomes. PointFinder (Zankari *et al.*, 2017) currently provides the databases for 10 bacterial taxa, including *E. coli*, *Salmonella enterica*, *Campylobacter jejuni*, *Mycobacterium tuberculosis*, *Enterococcus faecalis*, *Enterococcus faecium*, *Helicobacter pylori*, *Klebsiella* spp., *Neisseria gonorrhoeae*, and *Staphylococcus aureus*. The annotation algorithm used in PointFinder is based on the blastn search against known mutations. The RGI tool of the CARD provides an extensive list of known resistance mutations, currently housing 1,704 SNPs across 58 taxa (Alcock *et al.*, 2019). The AMRfinder currently covers seven taxa with a catalog of 682 point mutations and a detection module (Feldgarden *et al.*, 2019). ARG-ANNOT also provides the annotation of resistance mutations for *E. coli*, *M. tuberculosis*, and *Acinetobacter baumannii* (Gupta *et al.*, 2014), although it is not as extensive as the previously mentioned databases. In addition to databases covering multiple species, there are several resources specialized for the resistance mutations in single pathogen, such as *M. tuberculosis*.

WGS-based prediction of antibiotic resistance

The logic of phenotype prediction for antibiotic resistance differs from one method to another. The early type of predictors used predefined sets of known resistance determinants, such as the Mykrobe predictor (Bradley *et al.*, 2015). Prediction based on the presence of resistance determinants in the genome could be arguably the most straightforward method, but it has drawbacks mainly due to the risk of encountering novel resistance mechanisms and the incompleteness of our current knowledge on the genetic determinants of resistance. More recently, several studies used machine-learning approaches to address these problems. The input data used in the training and classification of machine-learning models can be the clustered catalogs of protein-coding genes (Van Camp *et al.*, 2020) or the genome-wide k-mers (Mahé and Tournoud, 2018). VAMPr provides 93 prediction models for nine well-studied bacterial species (Kim *et al.*, 2020). A highly accurate MIC prediction for non-typhoidal *Salmonella* has been demonstrated using the machine-learning model based on genomic k-mer features that are not restricted to known ARGs (Nguyen *et al.*, 2019). For *E. coli and Neisseria*, the pre-defined k-mers associated with resistance phenotypes were used to rapidly classify the phenotypic resistance of such strains present in samples (Břinda *et al.*, 2020). Interestingly, both promising results were based on genomic k-mer features that were not selected based on prior knowledge of ARG determinants.

According to systematic evaluation studies on the reliability of phenotypic resistance profiles (i.e., antibiograms) predicted from WGS data, predictions made by the current state-of-the-art methods are highly variable and lack reliability (Ellington *et al.*, 2017; Doyle *et al.*, 2020). While the methods for WGS-based prediction of antibiograms are still premature and unable to replace the traditional culture-dependent assays, the field is still growing with some promising developments made in recent years. As the increase in the available WGS data paired with clinical metadata is expected to aid the development of accurate models for the WGS-based prediction of resistance phenotypes, it is expected that the prediction capacity of WGS will be improved in the future.

Evolutionary and epidemiological analyses

For several research objectives, the interpretation of the WGS and phenotypic antibiotic resistance data of the given strain could be improved, when the WGS and the linked phenotypic resistance data (antibiogram) of closely related strains are available. Organized resources that meet such needs can be found at PATRIC (Antonopoulos *et al.*, 2019) and NDARO (Sayers *et al.*, 2020). It is also possible to search for NCBI BioSamples that have related antibiogram data using a query term including "antibiogram[filter]" to retrieve the list of samples that have the linked Sequence Read Archive data as well as the linked phenotypic data such as antibiotic susceptibility testing (AST) data (Kim *et al.*, 2020).

Interpretations regarding evolutionary or epidemiological circumstances can be derived from WGS data based on phylogenetic inferences. A typical workflow for whole-genome scale phylogenetic inference comprises three steps. First, the list of loci, either gene orthologs or single-nucleotide polymorphisms (SNPs), which are shared across all strains, are defined. This can be achieved using whole-genome multiple alignment tools such as progressiveMauve (Darling *et al.*, 2010) and Mugsy (Angiuoli and Salzberg, 2011), ortholog clustering methods such as Roary (Page *et al.*, 2015) and Pirate (Bayliss *et al.*, 2019), or the SNP calling tools that use a single reference genome such as Mummer (Marçais *et al.*, 2018) and Harvest suite (Treangen *et al.*, 2014). Once these loci are collected and aligned, the positions affected by homologous recombination can be removed using high-throughput screening tools such as Gubbins (Croucher *et al.*, 2015) and BratNextGen (Marttinen *et al.*, 2012) to improve the accuracy of phylogenetic inferences. Then, a maximum likelihood tree is usually constructed from the large alignment matrix using highly scalable tools such as FastTree (Price *et al.*, 2010) and IQ-Tree (Minh *et al.*, 2020). A phylogenetic tree can be used to estimate the timing of the divergences among the genomic lineages using tools such as BacDating (Didelot *et al.*, 2018) and BEAST (Bouckaert *et al.*, 2014). This type of approach can be used to investigate long-term scenarios of emergence of certain resistant variants emerge within the pathogenic species. A short-term epidemiological scenario (i.e., transmission path) of resistant clones could be tracked from the phylogeny using transmission tracking tools such as TransPhylo (Didelot *et al.*, 2017).

Antibiotic resistome analysis based on whole metagenome shotgun sequencing

Computational workflows for the cataloging of ARGs residing in the microbiomes using the metagenome shotgun sequencing data share several procedures in common with the WGS-based workflows described in the previous section. One of the key differences for the analysis of metagenome data is the objective of quantitative profiling in addition to the cataloging of the list of ARGs in samples. Another important difference is that for the sequences (i.e., raw reads or contigs) obtained from metagenome data generally cannot be decisively assigned with the origin of species. For this reason, the scope of the analysis is usually limited to the profiling of ARG families that can be identified based on homology alone, excluding the profiling of highly context-dependent resistance determinants such as point mutations.

Choice of workflows regarding the use of assembly step

The overall structures of computational approaches for metagenomic analysis of ARGs can be divided into three classes, based on the involvement of the assembly step. Analysis of sequence reads without assembling them offers quantitative profiles of the reference ARGs in samples. The resulting data are qualitatively equivalent to what can be achieved from highly multiplexed quantitative PCR assays, although the sequencing depth used limits the sensitivity of ARG detection. The use of assembly step in the workflow allows a wide range of downstream analyses, including the in-depth analysis of ARG sequences, taxonomic assignment of hosts carrying the ARGs, and investigation of MGE-ARG associations. One of the major concerns in the application of assembly-based approaches is the loss of information particularly when samples are from a complex microbiome and the sequencing data has low throughput. Gene-targeted assemblers provide the intermediate strategy, where the cataloging of full-length ARGs (i.e., as in the assembly-based approach) can be achieved with high sensitivity (i.e., as in the read-based approach).

Direct profiling of ARGs using metagenome sequencing reads

Using blastx or equivalent alignment tools, it is possible to directly profile the relative abundance of each reference ARG in samples using the unassembled raw reads as input. Because the relative abundance of ARGs represented by the number of reads aligned to each reference ARG sequence can be biased depending upon the length of each ARG and sequencing depth of each metagenome sample, appropriate normalization processes need to be done. To achieve normalized profiles (i.e., average copies per genome) of ARGs, two blastx runs are required against the ARG references and the standard marker genes, respectively. Standard marker genes need to be universal across bacteria, and 16S rRNA gene or a set of single-copy core genes can be used for this purpose. Blastx alignments against ARGs are normalized based on the length of each ARG and sequencing depth (e.g., reads per kb per million reads). Then, average per-genome copies of ARGs can be obtained by dividing the normalized reads of ARGs by those of universal single-copy genes or 16S rRNA gene. ARGs-OAP (Yin *et al.*, 2018) and DeepARG (Arango-Argoty *et al.*, 2018) are representatives of easily accessible computational pipelines serving this workflow. These pipelines also provide advantages by offering well-structured annotation of ARGs and adjusting blastx alignment cutoffs. However, many studies have been conducted with the manual, in-house pipelines possibly in pursuit of flexible database choices and high-throughput data processing in local computers. These read-based approaches provide higher sensitivity than the assembly-based approaches but have some restriction that no other information can be achieved than quantitative profiles of the reference ARGs.

Cataloguing ARGs by target protein-specific assemblers

The DNA or protein sequences of ARGs in metagenome samples are not determined in the read-based analysis. If a re-

search interest includes cataloging of the sequence variants present in the resistome of samples or the discovery of novel resistance genes, an assembly step is required. Protein-level assembly, as implemented in PLASS (Steinegger *et al*., 2019), and the target gene family-specific assemblers, such as fARGene (Berglund *et al*., 2019), can produce an order of magnitude more ORFs than the whole metagenome assembly approach, which will be described later. These tools are also computationally less demanding compared to the whole metagenome assembly. Moreover, this approach does not provide any linkage information on the host genomic context or taxonomy, and in the case of gene family-targeted assemblers, the assembly process should be repeated as many times as the number of gene families to be cataloged.

Cataloguing and quantitative profiling of ARGs using whole metagenome assembly

The assembly step demands the largest computational resources and time throughout the workflow, acting as a bottleneck in the assembly-based data analysis. The choices made for the tool and parameters in the assembly step can often be limited by the available computational resources. The two assemblers that are most popular in the metagenomics studies include Megahit (Li *et al*., 2015) and metaSPAdes (Nurk *et al*., 2017). The key parameter to adjust for performance optimization of these assemblers is the list of k-mer sizes to be used for de Bruijn graph construction.

Once the assembly is finished and the contigs are obtained, the next essential steps for the studies of antibiotic resistome are gene prediction and annotation of ARGs. Technically, performing these two steps with metagenome assemblies and draft genome assemblies does not differ. Therefore, the tools and databases used for WGS analysis are mostly applicable to metagenome analysis. Rapid blastp-like alignments are often used, as implemented in packages such as Diamond (Buchfink *et al*., 2015) and MMSeqs (Steinegger and Söding, 2017), instead of tools such as AMRfinder and RGI originally implemented with the target databases (e.g., AMRfinder and CARD) to speed up the analysis of large metagenomic data.

Once the ARG sequences have been catalogued from the contigs, in the following steps the ARGs are going to be quantitatively profiled and linked to the taxonomic assignments made at the contig-level. These two steps are unique aspects of metagenomic workflows. The gene-by-gene relative abundances within the metagenome of the sample can be calculated based on the coverage depth of the corresponding contig. The coverage depth of each contig, when not available in the assembler's output, should be calculated by mapping the raw reads against the contigs. Normalized abundance values that are suitable for inter-sample comparison can be calculated using the universal single-copy core genes as the standards. For assembly-based workflow, unlike in the case of raw read-based workflow, the use of 16S rRNA gene as the standard could be problematic because the 16S rRNA genes are known to be difficult to assemble. There are a number of metagenome taxonomic "profilers" available, which can estimate the taxonomic compositions from the input reads, such as MetaPhlAn (Beghini *et al*., 2020), but for the purpose of linking the ARGs to the taxonomy of the carrier bacteria, the right type of tools that can be used is "classifier", which assigns the taxon name to each input sequence. Centrifuge (Kim *et al*., 2016) and Kraken (Wood *et al*., 2019) are arguably the most popular metagenomic sequence classifiers in the field, both of which use k-mer exact matches between the query sequences and the taxonomy-linked reference genomes. However, these methods can lead to unambiguous taxonomic assignments at varying taxonomic ranks, from species to phylum, depending upon the length and sequence features of contigs. It is challenging to resolve taxonomy at low ranks (e.g., species or genus) when contigs are short or contain horizontally transferred sequences that are high homologous.

Identification of previously unknown ARGs and mobile resistome

Unknown ARGs

Considering how fast the emergence of novel ARGs has become a threat to antibiotic therapy that previously worked, there is undoubtedly an urgent need for cataloging the unknown ARGs present in diverse reservoirs. Such efforts expand the databases of the known ARGs and fill in the gaps between what could potentially emerge as a novel resistance in the clinic and what could be detected through WGS- or WMS-based surveillances.

Functional metagenomics has served as a key approach for culture-independent discovery of novel ARGs residing in diverse microbiomes. Generally, functional screening is performed on random fragments of metagenomic DNA in 1–5 kb of size, but it is also possible to construct a targeted library (e.g., integron gene cassettes) for functional screening (Böhm *et al*., 2020).

The databases and tools described in the previous sections are focused on the detection and annotation of the close homologs of ARGs that have been previously characterized in isolates, typically of well-characterized species. As a result, what can be cataloged as ARGs in the genomes and metagenomes are expected to be biased toward the branches of bacteria that have often been experimentally tested for antibiotic resistance. Indeed, the resistome of infant gut microbes cataloged by functional metagenomics shared a median identity of 32% with the reference protein sequences in the CARD (Gasparrini *et al*., 2019), emphasizing a wide gap between the true diversity and the known diversity of ARGs.

Searching for the distant homologs of known ARGs provides an alternative to find novel candidate ARG families. A number of recent studies have taken this approach to discover novel ARG families, including novel tetracycline resistance genes from various habitats (Berglund *et al*., 2020), novel colistin resistance genes from the gut microbiome of great apes (Campbell *et al*., 2020), and a total catalog of undescribed ARG families in the human gut metagenomes by aligning the sequences at the three-dimensional structure level (Ruppé *et al*., 2019). Tools developed in those studies, such as PCM (Ruppé *et al*., 2019) and fARGene (Berglund *et al*., 2019), can be employed in any projects sharing similar objectives.

It could also be useful to have additional ARG databases, instead of relying entirely on core ARG databases such as the CARD and AMRfinder. Putatively useful databases include

the Mustard database derived from the MetaHIT human gut microbiome protein catalogs (Ruppé *et al.*, 2019) and the FARME database containing sequences of metagenomic fragments that were functionally screened to confer antibiotic resistance (Wallace *et al.*, 2017).

Mobile resistome

The mobility of ARGs is considered to be correlated with the potential ability to spread rapidly across the globe as seen in the cases of *bla*KPC and *mcr* genes (Sheppard *et al.*, 2016; Wang *et al.*, 2018). The frequency of MGEs in the genomes varies substantially from one pathogen to another, suggesting that the contribution of the mobile ARGs to the overall antibiotic resistance development likely varies across species. Indeed, the known resistance determinants are dominantly point mutations in some species such as *M. tuberculosis* and *N. gonnorhoeae*, whereas acquired mobile genes comprise the majority of known resistance determinants in some pathogens such as those in *Enterobacteriaceae* species (Ellington *et al.*, 2017). For several reasons, it is valuable to assess the mobility of ARGs detected in the genomes and metagenomes.

We propose the definition of mobile resistome in a broad sense to include the following: (a) the ARGs located within the MGEs, thereby indicating mobilization in the recent past and prone to future mobilization and HGT (the MGE-borne ARGs), and (b) the ARGs whose phyletic distribution shows the signatures of previous mobilization across bacterial lineages, namely HGT, irrespective of their current linkage to the MGE context (the mobilized ARGs). In accordance with the broad definition of mobile resistome that we explained above, the approaches for identifying mobile resistome can be employed in two ways.

One group of approaches is targeted at the identification of MGE-borne ARGs. Of the known forms of MGEs, plasmids, integrative and conjugative elements (ICEs), insertion sequences (ISs), transposons, integrons, and phages could be named as the most extensively characterized MGEs for their involvement in the mobilization of ARGs among bacteria. Computational tools can be used to screen for MGEs in genomic or metagenomic sequences, such as MOB-suite and

Table 1. Useful tools and databases for the sequencing-based studies on antibiotic resistance

Step	Tool/Database	Note & Website	Applied to (mostly)
Pre-processing and assembly	Fastp (Chen *et al.*, 2018)	Short-read pre-processing and quality control github.com/OpenGene/fastp	WGS, WMS
	SPAdes/MetaSPAdes (Bankevich *et al.*, 2012; Nurk *et al.*, 2017)	Short-read WGS/WMS assembly cab.spbu.ru/software/spades	WGS
	Flye/metaFlye (Kolmogorov *et al.*, 2019, 2020)	Long-read WGS/WMS assembly github.com/fenderglass/Flye	WGS, WMS
	Megahit (Li *et al.*, 2015)	Short-read WMS assembly github.com/voutcn/megahit	WMS
	PLASS (Steinegger *et al.*, 2019)	Protein-level WMS assembly github.com/soedinglab/plass	WMS
Species identification	GTDB/GTDB-Tk (Parks *et al.*, 2018; Chaumeil *et al.*, 2019)	Curated database for genome taxonomy with the standalone tool for species ID for genomes with strength in uncultured prokaryotes gtdb.ecogenomic.org	WGS
	EzBioCloud (Yoon *et al.*, 2017)	Taxonomy-curated 16S rRNA genes and genomes www.ezbiocloud.net	WGS
Strain subtyping	PubMLST (Jolley *et al.*, 2018)	Collection of publicly available MLST schemes pubmlst.org	WGS
	MLSTcheck (Page *et al.*, 2016)	MLST assignment on the input genome assembly using PubMLST schemes github.com/sanger-pathogens/mlst_check	WGS
	PopPUNK (Lees *et al.*, 2019)	Strain clustering based on whole genome k-mers github.com/johnlees/PopPUNK	WGS
ARGs	CARD/RGI (Alcock *et al.*, 2019)	Manually curated database of experimentally validated ARGs and the annotation tool applicable to assembled sequences card.mcmaster.ca/	WGS, WMS
	AMRFinder (Feldgarden *et al.*, 2019)	ARG database curated by NCBI and the annotation tool applicable to assembled sequences ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder	WGS, WMS
	ResFinder/PointFinder (Zankari *et al.*, 2017; Bortolaia *et al.*, 2020)	Tools for identification of acquired ARGs and point mutations based on the databases curated by CGE cge.cbs.dtu.dk/services/ResFinder	WGS, WMS
	ARGs-OAP/SARGs (Yin *et al.*, 2018)	Direct profiling of ARGs from metagenome reads smile.hku.hk/SARGs	WMS
	DeepARG (Arango-Argoty *et al.*, 2018)	Direct profiling of ARGs from metagenome reads bench.cs.vt.edu/deeparg	WMS
Inter1pretation	NDARO (Sayers *et al.*, 2020)	Database of antibiotic resistant organisms including the antibiograms and genome sequences ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance	WGS
	PATRIC (Antonopoulos *et al.*, 2019)	Database of pathogens including phenotypic resistance and genome sequences www.patricbrc.org	WGS

ARG, antibiotic resistance gene; WGS, whole genome sequencing; WMS, whole metagenome sequencing.

Platon for plasmids (Robertson and Nash, 2018; Schwengers *et al.*, 2020), I-VIP for integrons (Zhang *et al.*, 2018), ISEScan (Xie and Tang, 2017) for conjugative transposons, and ICEfinder for ICEs (Liu *et al.*, 2019). Alternatively, targeted metagenomic sequencing strategies have been used to achieve deep sequencing of the plasmidome (Kothari *et al.*, 2019), integron cassettes (Ghaly *et al.*, 2019), extracellular DNA (Yuan *et al.*, 2019), and associated resistomes. From shotgun metagenomes, spacegraphcats (Brown *et al.*, 2020), and MetaCherchant (Olekhnovich *et al.*, 2018) can be used to extract the ARG's neighboring contexts over complex assembly graphs to search for MGE signatures.

The other group of approaches utilizes the evidence of HGT indicated by either phylogenomic analysis or innovative sequencing strategies that produce linkage information about bacterial hosts carrying the ARGs. Comparison of the genomes of closely related strains can reveal the genomic islands that are present only in a few strains, indicating the horizontal acquisition of the genomic region. Phylogenomic analysis can reveal clusters of almost identical (or highly similar) genes shared across distant bacteria (e.g., different species) (Smillie *et al.*, 2011). This logic has been applied to a variety of data sets to characterize the shared mobile resistome across taxa and ecosystems (Lee *et al.*, 2020). There are also recently developed innovative sequencing techniques, such as Hi-C (Kent *et al.*, 2020), epic-PCR (Spencer *et al.*, 2016), and sequencing of CRISPR-Cas spacer sequences (Munck *et al.*, 2020), which can provide evidence of transfer and acquisition events in real time when applied to time-series data sets.

A recent study revealed that chromosomal locations of horizontally transferred genes often lack MGEs (Oliveira *et al.*, 2017). It has also been suggested that the structural parts of MGEs can be lost rapidly in the recipient's genome through pervasive deletion and recombination (Smillie *et al.*, 2011; Brito *et al.*, 2016). Furthermore, MGE-borne ARGs may not always show evidence of HGTs in the available genome dataset. Therefore, the evidence of HGT and the presence of MGE structure are complementary to each other as the means of identifying mobile reistome and are often sought one after another to corroborate the mobile nature of the ARGs.

Conclusion

There is a wealth of data analysis tools available for WGS- or WMS-based cataloging of known ARGs in the bacterial isolates or communities (see Table 1 for the short list of useful tools and databases). Methodological innovations are still ongoing in the exploration of mobile resistome, and advancements in sequencing technologies such as long-read sequencing and Hi-C libraries, which facilitate new approaches to explore mobilome-resistome connections. Sequencing-based prediction of phenotypic resistance remains relatively immature at present, but the accumulation of genome data linked to phenotypic data will improve the overall situation. Considering a variety of analytical tools and newly developed rapid methods, it is critical to properly understand and select appropriate tools and methods when designing new studies on the antibiotic resistome and mobilome.

Acknowledgments

This work was supported by the Korea Ministry of Environment (MOE) as "the Environmental Health Action Program (2016001350004)" and by the "Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ015130032020)" Rural Development Administration, Republic of Korea.

Conflict of Interest

The authors declare that they have no conflict of interest

References

Alcock, B.P., Raphenya, A.R., Lau, T.T.Y., Tsang, K.K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.L.V., Cheng, A.A., Liu, S., *et al.* 2019. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525.

Angiuoli, S.V. and Salzberg, S.L. 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27**, 334–342.

Antonopoulos, D.A., Assaf, R., Aziz, R.K., Brettin, T., Bun, C., Conrad, N., Davis, J.J., Dietrich, E.M., Disz, T., Gerdes, S., *et al.* 2019. PATRIC as a unique resource for studying antimicrobial resistance. *Brief. Bioinform.* **20**, 1094–1102.

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L.S., Vikesland, P., and Zhang, L. 2018. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., *et al.* 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477.

Bayliss, S.C., Thorpe, H.A., Coyle, N.M., Sheppard, S.K., and Feil, E.J. 2019. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *GigaScience* **8**, giz119.

Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Thomas, A.M., Manghi, P., Valles-Colomer, M., *et al.* 2020. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *bioRxiv*. doi: https://doi.org/10.1101/2020.11.19.388223

Berglund, F., Böhm, M.E., Martinsson, A., Ebmeyer, S., Österlund, T., Johnning, A., Larsson, D.G.J., and Kristiansson, E. 2020. Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes. *Microb. Genom.* **6**, mgen000455.

Berglund, F., Österlund, T., Boulund, F., Marathe, N.P., Larsson, D.G.J., and Kristiansson, E. 2019. Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome* **7**, 52.

Böhm, M.E., Razavi, M., Marathe, N.P., Flach, C.F., and Larsson, D.G.J. 2020. Discovery of a novel integron-borne aminoglycoside resistance gene present in clinical pathogens by screening environmental bacterial communities. *Microbiome* **8**, 41.

Bortolaia, V., Kaas, R.S., Ruppe, E., Roberts, M.C., Schwarz, S., Cattoir, V., Philippon, A., Allesoe, R.L., Rebelo, A.R., Florensa, A.F., *et al.* 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **75**, 3491–3500.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D.,

Suchard, M.A., Rambaut, A., and Drummond, A.J. 2014. BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537.

Bradley, P., Gordon, N.C., Walker, T.M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L.J., Anson, L., de Cesare, M., *et al.* 2015. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat. Commun.* **6**, 10063.

Břinda, K., Callendrello, A., Ma, K.C., MacFadden, D.R., Charalampous, T., Lee, R.S., Cowley, L., Wadsworth, C.B., Grad, Y.H., Kucherov, G., *et al.* 2020. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nat. Microbiol.* **5**, 455–464.

Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W., Tamminen, M., Smillie, C.S., Wortman, J.R., *et al.* 2016. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439.

Brown, C.T., Moritz, D., O'Brien, M.P., Reidl, F., Reiter, T., and Sullivan, B.D. 2020. Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity. *Genome Biol.* **21**, 164.

Buchfink, B., Xie, C., and Huson, D.H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60.

Campbell, T.P., Sun, X., Patel, V.H., Sanz, C., Morgan, D., and Dantas, G. 2020. The microbiome and resistome of chimpanzees, gorillas, and humans across host lifestyle and geography. *ISME J.* **14**, 1584–1599.

Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890.

Cohen, K.A., Manson, A.L., Desjardins, C.A., Abeel, T., and Earl, A.M. 2019. Deciphering drug resistance in *Mycobacterium tuberculosis* using whole-genome sequencing: progress, promise, and challenges. *Genome Med.* **11**, 45.

Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J., and Harris, S.R. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15.

Darling, A.E., Mau, B., and Perna, N.T. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147.

Didelot, X., Croucher, N.J., Bentley, S.D., Harris, S.R., and Wilson, D.J. 2018. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134.

Didelot, X., Fraser, C., Gardy, J., and Colijn, C. 2017. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007.

Doyle, R.M., O'Sullivan, D.M., Aller, S.D., Bruchmann, S., Clark, T., Coello Pelegrin, A., Cormican, M., Diez Benavente, E., Ellington, M.J., McGrath, E., *et al.* 2020. Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. *Microb. Genom.* **6**, e000335.

Ellington, M.J., Ekelund, O., Aarestrup, F.M., Canton, R., Doumith, M., Giske, C., Grundman, H., Hasman, H., Holden, M.T.G., Hopkins, K.L., *et al.* 2017. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin. Microbiol. Infect.* **23**, 2–22.

Feldgarden, M., Brover, V., Haft, D.H., Prasad, A.B., Slotta, D.J., Tolstoy, I., Tyson, G.H., Zhao, S., Hsu, C.H., McDermott, P.F., *et al.* 2019. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* **63**, e00483.

Gasparrini, A.J., Wang, B., Sun, X., Kennedy, E.A., Hernandez-Leyva, A., Ndao, I.M., Tarr, P.I., Warner, B.B., and Dantas, G. 2019. Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nat. Microbiol.* **4**, 2285–2297.

Ghaly, T.M., Geoghegan, J.L., Alroy, J., and Gillings, M.R. 2019. High diversity and rapid spatial turnover of integron gene cassettes in soil. *Environ. Microbiol.* **21**, 1567–1574.

Gupta, S.K., Padmanabhan, B.R., Diene, S.M., Lopez-Rojas, R., Kempf, M., Landraud, L., and Rolain, J.M. 2014. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.* **58**, 212–220.

Joensen, K.G., Tetzschner, A.M.M., Iguchi, A., Aarestrup, F.M., and Scheutz, F. 2015. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J. Clin. Microbiol.* **53**, 2410–2426.

Jolley, K.A., Bray, J.E., and Maiden, M.C.J. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **3**, 124.

Kent, A.G., Vill, A.C., Shi, Q., Satlin, M.J., and Brito, I.L. 2020. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nat. Commun.* **11**, 4379.

Kim, J., Greenberg, D.E., Pifer, R., Jiang, S., Xiao, G., Shelburne, S.A., Koh, A., Xie, Y., and Zhan, X. 2020. VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. *PLoS Comput. Biol.* **16**, e1007511.

Kim, D., Song, L., Breitwieser, F.P., and Salzberg, S.L. 2016. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729.

Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P.L., *et al.* 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546.

Kothari, A., Wu, Y.W., Chandonia, J.M., Charrier, M., Rajeev, L., Rocha, A.M., Joyner, D.C., Hazen, T.C., Singer, S.W., and Mukhopadhyay, A. 2019. Large circular plasmids from groundwater plasmidomes span multiple incompatibility groups and are enriched in multimetal resistance genes. *mBio* **10**, e02899-18.

Lee, K., Kim, D.W., Lee, D.H., Kim, Y.S., Bu, J.H., Cha, J.H., Thawng, C.N., Hwang, E.M., Seong, H.J., Sul, W.J., *et al.* 2020. Mobile resistome of human gut and pathogen drives anthropogenic bloom of antibiotic resistance. *Microbiome* **8**, 2.

Lees, J.A., Harris, S.R., Tonkin-Hill, G., Gladstone, R.A., Lo, S.W., Weiser, J.N., Corander, J., Bentley, S.D., and Croucher, N.J. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316.

Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676.

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z., and Ou, H.Y. 2019. ICEberg 2.0: An updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.* **47**, D660–D665.

MacFadden, D.R., Coburn, B., Břinda, K., Corbeil, A., Daneman, N., Fisman, D., Lee, R.S., Lipsitch, M., McGeer, A., Melano, R.G., *et al.* 2020. Using genetic distance from archived samples for the prediction of antibiotic resistance in *Escherichia coli*. *Antimicrob. Agents Chemother.* **64**, e02417-19.

Mahé, P. and Tournoud, M. 2018. Predicting bacterial resistance from whole-genome sequences using *k*-mers and stability selection. *BMC Bioinformatics* **19**, 383.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. 2018. MUMmer4: A fast and versatile genome align-

ment system. *PLoS Comput. Biol.* **14**, e1005944.
Marttinen, P., Hanage, W.P., Croucher, N.J., Connor, T.R., Harris, S.R., Bentley, S.D., and Corander, J. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res*. **40**, e6.
Meier-Kolthoff, J.P. and Göker, M. 2019. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat. Commun*. **10**, 2182.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534.
Munck, C., Sheth, R.U., Freedberg, D.E., and Wang, H.H. 2020. Recording mobile DNA in the gut microbiota using an *Escherichia coli* CRISPR-Cas spacer acquisition platform. *Nat. Commun.* **11**, 95.
Nguyen, M., Long, S.W., McDermott, P.F., Olsen, R.J., Olson, R., Stevens, R.L., Tyson, G.H., Zhao, S., and Davis, J.J. 2019. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.* **57**, e01260.
Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. 2017. metaSPAdes: A new versatile metagenomic assembler. *Genome Res*. **27**, 824–834.
Olekhnovich, E.I., Vasilyev, A.T., Ulyantsev, V.I., Kostryukova, E.S., and Tyakht, A.V. 2018. MetaCherchant: analyzing genomic context of antibiotic resistance genes in gut microbiota. *Bioinformatics* **34**, 434–444.
Oliveira, P.H., Touchon, M., Cury, J., and Rocha, E.P.C. 2017. The chromosomal organization of horizontal gene transfer in bacteria. *Nat. Commun.* **8**, 841.
Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132.
Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. 2015. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693.
Page, A.J., Taylor, B., and Keane, J.A. 2016. Multilocus sequence typing by blast from *de novo* assemblies against PubMLST. *J. Open Source Softw.* **1**, 118.
Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., and Hugenholtz, P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004.
Perez-Sepulveda, B.M., Heavens, D., Pulford, C.V., Predeus, A.V., Low, R., Webster, H., Schudoma, C., Rowe, W., Lipscombe, J., Watkins, C., *et al.* 2020. An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. *bioRxiv* 200840. doi: https://doi.org/10.1101/2020.07.22.200840.
Price, M.N., Dehal, P.S., and Arkin, A.P. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490.
Robertson, J. and Nash, J.H.E. 2018. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom*. **4**, e000206.
Ruppé, E., Ghozlane, A., Tap, J., Pons, N., Alvarez, A.S., Maziers, N., Cuesta, T., Hernando-Amado, S., Clares, I., Martínez, J.L., *et al.* 2019. Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nat. Microbiol.* **4**, 112–123.
Sayers, E.W., Beck, J., Brister, J.R., Bolton, E.E., Canese, K., Comeau, D.C., Funk, K., Ketter, A., Kim, S., Kimchi, A., *et al.* 2020. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. **48**, D9–D16.
Schwengers, O., Barth, P., Falgenhauer, L., Hain, T., Chakraborty, T., and Goesmann, A. 2020. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb. Genom.* **6**, mgen000398.
Segerman, B. 2020. The most frequently used sequencing technologies and assembly methods in different time segments of the bacterial surveillance and RefSeq genome databases. *Front. Cell. Infect. Microbiol*. **10**, 527102.
Sheppard, A.E., Stoesser, N., Wilson, D.J., Sebra, R., Kasarskis, A., Anson, L.W., Giess, A., Pankhurst, L.J., Vaughan, A., Grim, C.J., *et al.* 2016. Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene bla_{KPC}. *Antimicrob. Agents Chemother*. **60**, 3767–3778.
Shi, W., Sun, Q., Fan, G., Hideaki, S., Moriya, O., Itoh, T., Zhou, Y., Cai, M., Kim, S.G., Lee, J.S., *et al.* 2020. gcType: a high-quality type strain genome database for microbial phylogenetic and functional research. *Nucleic Acids Res*. **49**, D694–D705.
Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., and Alm, E.J. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244.
Spencer, S.J., Tamminen, M.V., Preheim, S.P., Guo, M.T., Briggs, A.W., Brito, I.L., Weitz, D.A., Pitkänen, L.K., Vigneault, F., Virta, M.P.J., *et al.* 2016. Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *ISME J.* **10**, 427–436.
Steinegger, M., Mirdita, M., and Söding, J. 2019. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* **16**, 603–606.
Steinegger, M. and Söding, J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**. 1026–1028.
Treangen, T.J., Ondov, B.D., Koren, S., and Phillippy, A.M. 2014. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **15**, 524.
Van Camp, P.J., Haslam, D.B., and Porollo, A. 2020. Prediction of antimicrobial resistance in Gram-negative bacteria from whole-genome sequencing data. *Front. Microbiol.* **11**, 1013.
Wallace, J.C., Port, J.A., Smith, M.N., and Faustman, E.M. 2017. FARME DB: a functional antibiotic resistance element database. *Database* **2017**, baw165.
Wang, R., van Dorp, L., Shaw, L.P., Bradley, P., Wang, Q., Wang, X., Jin, L., Zhang, Q., Liu, Y., Rieux, A., *et al.* 2018. The global distribution and spread of the mobilized colistin resistance gene *mcr-1*. *Nat. Commun.* **9**, 1179.
Wick, R.R., Heinz, E., Holt, K.E., and Wyres, K.L. 2018. Kaptive Web: user-friendly capsule and lipopolysaccharide serotype prediction for *Klebsiella* genomes. *J. Clin. Microbiol.* **56**, e00197-18.
Wick, R.R. and Holt, K.E. 2020. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res*. **8**, 2138.
Wood, D.E., Lu, J., and Langmead, B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257.
Xie, Z. and Tang, H. 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**, 3340–3347.
Yin, X., Jiang, X.T., Chai, B., Li, L., Yang, Y., Cole, J.R., Tiedje, J.M., and Zhang, T. 2018. 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* **34**, 2263–2270.
Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H., and Chun, J. 2017. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* **67**, 1613–1617.
Yoshida, C.E., Kruczkiewicz, P., Laing, C.R., Lingohr, E.J., Gannon, V.P.J., Nash, J.H.E., and Taboada, E.N. 2016. The *Salmonella in silico* typing resource (SISTR): an open web-accessible tool for

rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS ONE* **11**, e0147101.

Yuan, Q.B., Huang, Y.M., Wu, W.B., Zuo, P., Hu, N., Zhou, Y.Z., and Alvarez, P.J.J. 2019. Redistribution of intracellular and extracellular free & adsorbed antibiotic resistance genes through a wastewater treatment plant by an enhanced extracellular DNA extraction method with magnetic beads. *Environ. Int.* **131**, 104986.

Zankari, E., Allesøe, R., Joensen, K.G., Cavaco, L.M., Lund, O., and Aarestrup, F.M. 2017. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J. Antimicrob. Chemother.* **72**, 2764–2768.

Zhang, A.N., Li, L.G., Ma, L., Gillings, M.R., Tiedje, J.M., and Zhang T. 2018. Conserved phylogenetic distribution and limited antibiotic resistance of class 1 integrons revealed by assessing the bacterial genome and plasmid collection. *Microbiome* **6**, 130.

Zhou, Z., Alikhan, N.F., Mohamed, K., Fan, Y., Agama Study Group, and Achtman, M. 2019. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res.* **30**, 138–152.