CARPDM: cost-effective antibiotic resistome profiling of metagenomic samples using targeted enrichment

Dirk Hackenberger,[1,2] Hamna Imtiaz,[1,2] Amogelang R. Raphenya,[1,2] Brian P. Alcock,[1,2] Hendrik N. Poinar,[1,2,3,4,5] Gerard D. Wright,[1,2] Andrew G. McArthur[1,2]

AUTHOR AFFILIATIONS See affiliation list on p. 16.

ABSTRACT   Better interrogation of antimicrobial resistance requires new approaches to detect the associated genes in metagenomic samples. Targeted enrichment is an ideal method for their sequencing and characterization. However, no open-source, up-to-date hybridization probe set targeting antimicrobial resistance genes exists. Here, we describe the Comprehensive Antibiotic Resistance Probe Design Machine (CARPDM), a probe design software package made to run alongside all future Comprehensive Antibiotic Resistance Database releases. To test its efficacy, we have created and tested two separate probe sets: allCARD, which enriches all genes encoded in the Comprehensive Antibiotic Resistance Database's protein homolog models ($n = 4,661$), and clinicalCARD, which focuses on a clinically relevant subset of resistance genes ($n = 323$). We demonstrate that allCARD increases the number of reads mapping to resistance genes by up to 594-fold. clinicalCARD performs similarly when clinically relevant genes are present, increasing the number of resistance-gene mapping reads by up to 598-fold. In parallel with this development, we have established a protocol to synthesize any probe set in-house, saving up to 350 dollars per reaction. Together, these probe sets, their associated design program CARPDM, and the protocol for in-house synthesis will democratize metagenomic resistome analyses, allowing researchers access to a cost-effective and efficient means to explore the antibiotic resistome.

IMPORTANCE   Antimicrobial resistance threatens to undermine all modern medicine and is driven by the spread of antimicrobial resistance genes among pathogens, environments, patients, and animals. DNA sequencing of complex samples, such as wastewater, shows considerable promise for tracking these genes and making risk assessments. However, these methods suffer from high costs and low detection limits, plus a requirement for frequent redesign due to the constantly evolving diversity of resistance genes. Building upon our Comprehensive Antibiotic Resistance Database, our research provides software for on-demand renewal, based on the latest knowledge of resistance gene diversity, of our novel bait-capture hybridization platform that simultaneously reduces cost and increases detection levels for DNA sequencing of complex samples. The significance of our research is in the development of new software tools, reagent synthesis protocols, and hybridization enrichment protocols to provide affordable, high-resolution metagenomics DNA sequencing, which we test using environmental and wastewater samples.

KEYWORDS   antimicrobial resistance, sequencing, metagenomics, target enrichment

Antimicrobial resistance (AMR) is a growing and global problem. In 2019, AMR was estimated to be directly responsible for 1.27 M deaths (1). By 2050, this number may be as high as 10 M (2). Most of this impact is and will continue to be in regions least equipped to combat it, largely due to a lack of resources (1). Therefore, it is imperative

that we devise cost-effective solutions to address AMR more effectively and equitably. Bacterial evolution has been marked by the arms race of antibiotics and their associated resistance genes, providing a competitive edge to their producers (3–8). This struggle is revealed in the vast reservoir of antimicrobial resistance genes (ARGs) in environmental microbes, requiring only mobilization through horizontal gene transfer to be effective against any treatment we deploy (9–12). Therefore, to better combat AMR emergence, we need to improve how we detect the full complement of ARGs (i.e., the resistome [13]) in environmental and other reservoirs. Attractive monitoring targets are environments such as wastewater, a fertile ground for genetic exchange between bacteria (5) that provides a snapshot of ARG prevalence in the community (14–20). Natural environments, such as soils, rivers, and farms, are known reservoirs of ARGs, many with the potential to mobilize into pathogens (21–25). Finally, profiling the resistome of human and animal microbiomes allows us to identify critical determinants in the spread of resistance within and between these two groups (26, 27).

Investigating these rich data sources requires methods to characterize their resistome. Several techniques exist that may fill this niche, each with its limitations. PCR, for example, is commonly used to detect ARGs in *Mycobacterium tuberculosis* isolates (28). While useful for a small, targeted set of genes, the Comprehensive Antibiotic Resistance Database (CARD) (29) hosts over 5,000 resistance determinants, an untenable number for PCR methods. Furthermore, because of the specificity of PCR, there is little chance of detecting distantly related genes, as a single nucleotide substitution may eliminate any signal from the assay. Finally, even if one detects a novel sequence variant by PCR, without follow-up amplicon sequencing, there is no way to identify it. This makes phylogenetic tracking of the spread of AMR far more difficult.

Shotgun DNA sequencing can detect all the genes in a sample given sufficient depth. Groups have used this method to characterize environmental (25, 30) and worldwide wastewater (14, 15) resistomes. A limitation of this technique is that all ARGs in a metagenomic sample typically represent <1% of the total DNA, even in samples where they are highly abundant, such as wastewater. Individual ARGs may be several orders of magnitude less than that in abundance. For example, a single 1 kb ARG that makes up 1×10^{-6}% of the DNA in a sample would require 10 Gbp of sequence data to obtain 10-fold coverage of the ARG. Performing this work on a NextSeq 2,000 would cost over USD$1,500. As such, while deep sequencing can be used to characterize metagenomic resistomes, it entails a high cost per sample, most of which will be spent sequencing background DNA. The associated volume of data also increases equipment costs and computing power needed to parse the data, further constraining this technique's use in resource-limited settings (31).

Targeted enrichment is a modification to shotgun sequencing that allows robust detection of a broad range of specific, low-abundance targets with less sequencing. In this protocol, DNA from a sequencing library is denatured, allowing biotinylated RNA "probes" complementary to a set of target sequences to hybridize (32, 33). Streptavidin-coated magnetic beads capture these biotinylated RNA probes and their complementary DNA partners from the background (Fig. 1). This process increases the proportion of the target DNA in a library, allowing one to sequence less yet detect more.

A probe-capture protocol is ideal for detecting thousands of ARGs with a fraction of the sequencing required by brute-force shotgun approaches. However, two challenges exist for the probe-capture strategy in the AMR space. First, there is no up-to-date and open-source AMR probe set, as the most recent was designed against only 2,021 ARGs from CARD v1.0.1, released in 2015 (34). The second is that the cost of probes from commercial suppliers can be up to $350 per reaction, diminishing cost savings relative to shotgun sequencing. To address the first of these challenges, we have written a software package—the Comprehensive Antibiotic Resistance Probe Design Machine (CARPDM)—to generate a stringently filtered probe set with minimal off-target enrichment from the CARD v3.2.5 protein homolog model (allCARD) ARGs. This software package will run alongside all new releases of CARD, ensuring there is always an open-source and
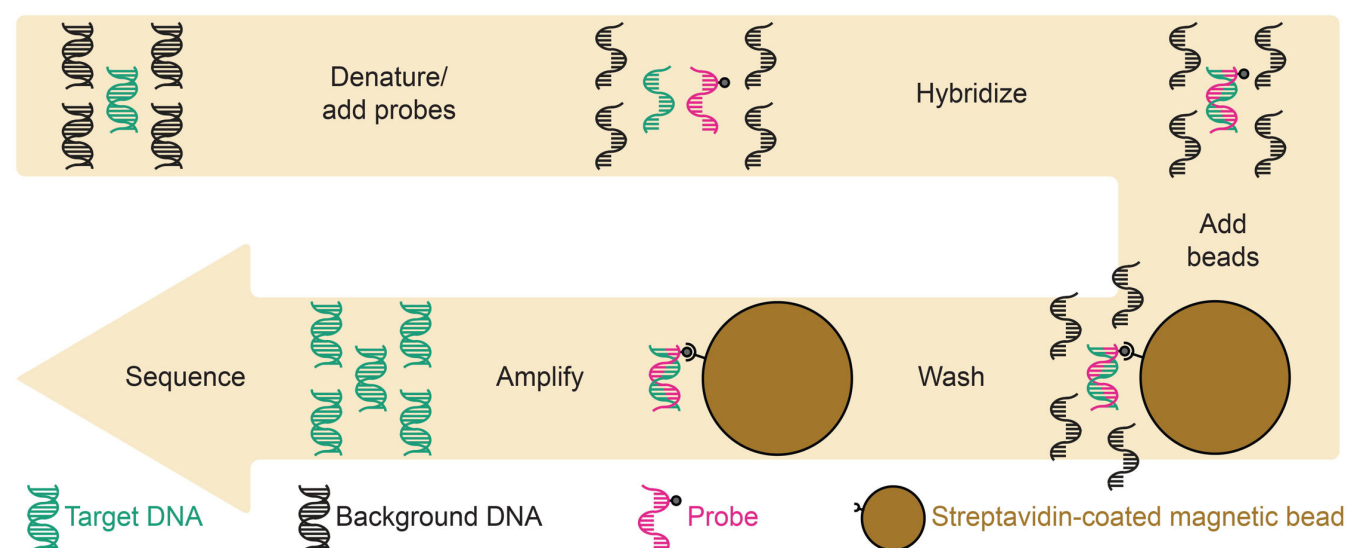
FIG 1 Targeted enrichment workflow. Magnetic streptavidin beads are used to bind biotinylated RNA molecules, which are, in turn, attached to a complementary DNA partner. One can considerably bias the resulting library toward a target fraction by pelleting the beads and washing away the background before amplification and sequencing.

up-to-date probe set to enrich ARGs from any sample. We also curated a list of 323 clinically relevant ARGs and generated a smaller probe set (clinicalCARD) with the same program, providing a more focused alternative to the allCARD probe set, as the full complement of ARGs may not be necessary for many projects in healthcare settings. To address the cost of commercial probes, we have developed a protocol that allows in-house synthesis of any probe set from a Twist Biosciences oligo-pool. With this strategy, researchers can synthesize thousands of reactions worth of any probe set for a one-time fee lower than that of a typical 16-reaction kit from a commercial supplier.

To test our probe sets and in-house synthesis of probes, five conditions were tested on two wastewater and three soil samples in duplicate, comparing enrichment via the 2015 CARD v1.0.1 (34), allCARD, clinicalCARD probe sets to sequencing without enrichment, plus examination of the comparative performance of commercially supplied probes, and probes synthesized using our new protocol. We demonstrate that our in-house synthesized probe set detected more ARGs with the same amount of sequencing on the samples we tested when compared to a commonly used commercial option. We also illustrate CARPDM's capability as a probe design platform by testing allCARD and clinicalCARD probe sets on wastewater and soil samples. We found that allCARD detects far more ARGs than the probe set generated against CARD v1.0.1 (34). Finally, we show that clinicalCARD detects clinically relevant ARGs with less sequencing and better coverage than allCARD.

MATERIALS AND METHODS

ARG selection for allCARD

Only ARGs curated as protein homolog models were included from CARD v3.2.5 (n = 4,661), as these ARGs do not confer resistance via the acquisition of mutations (i.e., CARD's protein variant models). Including mutation-based ARGs in the design would enrich wild-type alleles, as enrichment probes can readily hybridize over a single mutation, diluting the sequencing effort.

ARG selection for clinicalCARD

The CARD prevalence data v4.0.0 (19) aided in identifying clinically relevant ARGs. This data set is a collection of 221,175 sequencing assemblies from 377 pathogens with associated ARGs identified by CARD's Resistome Gene Identifier (RGI) software (35). These data include a collection of 21,079 completely sequenced chromosomes and 41,828 completely sequenced plasmids. Preliminary examination of these data illustrated that the distribution of ARGs was almost binary, i.e., each ARG predominantly occurred in either plasmids or chromosomes, but not both (Fig. S1). It also showed that ARGs with a higher occurrence of plasmids were far more likely to be clinically relevant (36–39). As such, the first list of candidate clinically relevant ARGs included any ARG in CARD with at least 10 occurrences in plasmids, yielding 237 ARGs. Ten was chosen as a cutoff as it minimized the false positive inclusion of highly prevalent but clinically irrelevant ARGs such as species-specific efflux mechanisms. Any ARG with >5% but <95% prevalence in any ESKAPE pathogen (40, 41) (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* species) was also added to this list, yielding a further 124 ARGs. Finally, 42 ARGs identified as clinically significant via a literature review were added, as they were included in publications investigating similar questions (42, 43). Overall, a wide net was cast, erring on the side of false positives and resulting in a draft set of 403 candidate ARGs. This set was manually curated based on CARD prevalence data, considering the species they appeared in, the genomic context (whether it was mobile, i.e., on a plasmid), the drug classes they impact, the environmental overlap of the host organism with humans, and the relative risk each host imposed as a pathogen. Our approach was developed following recent ARG risk frameworks developed for metagenomic data (42, 43). In short, our manual curation selected for mobile ARGs that confer resistance to antibiotics used for therapy in humans is present in a human pathogen and is likely to colonize humans or animals. After curation, 323 ARGs were deemed clinically relevant and included in the clinicalCARD probe set design.

Probe design

CARPDM is written in Python, save for the first step, which employs the program BaitsTools (44) to tile probe sequences along ARG sequences. Probes designed by CARPDM have an 80 nt length as per the prior validated ARG bait set for CARD v1.0.1. This is also an ideal size for adding amplification primers while remaining below the 120 nt cut-off for the base level of a Twist oligo-pool if synthesizing probes in-house. The probes also have an extremely high density before filtering, with a tiling distance of 4 nt along ARG nucleotide sequences. This increases redundancy in the probe set and makes it more robust against stringent downstream filtering. As input, the allCARD design used all 4,661 CARD v3.2.5 protein homolog model ARGs, while clinicalCARD's design used the curated 323 clinically relevant ARGs.

There are three filtering steps after the initial probe construction (Fig. S2). The first is a filter based on sequence, in which the probes are deduplicated, confirmed as 80 nt, contain no ambiguous bases, have no perfect complements within the set, have a $T_m > 50°C$, and contain no LguI restriction endonuclease cut sites. The last condition is essential for the synthesis protocol. Since bacteria are likely to be the most abundant organism in most metagenomic samples (45), the second filter is a BLASTN (46) search against the nt database to minimize off-target enrichment and remove probes with >80% identity over <50 nt to any bacterial sequence, as any probe similar to a bacterial sequence over >50 nt is likely the ARG itself. This filter also removes probes with >80% similarity over >50 nt to any viral, archaeal, or eukaryotic sequence unless they match perfectly (i.e., likely sequence contamination by an ARG).

The last filter removes redundancy in the probe set, using BLASTN to compare the probes against each other. To remove probes that could bind to each other rather than the target DNA, any probes found to be complementary over the entire 80 nt are collapsed to a single representative. After removing complementary probes, CARPDM

determines if the number of probes in the set is below a pre-defined cut-off. For allCARD, we arbitrarily set this cut-off at 40,000 probes, while for clinicalCARD, it was set to 20,000 probes. If not below the cut-off, CARPDM collapses probes with 79 nt of complementarity to a single representative, repeating the process with decreasing length of complementarity until below the cut-off.

In silico analysis of probe sets

BLASTN (46) was used to compare the probe sets to the ARG sequences. Using Python and NumPy (47), "target" arrays of zeros were constructed. Each of these arrays represented a single ARG against which the probeset was designed. After aligning all probes against the input set of target sequences with BLASTN, arrays of ones (matches) and zeros (mismatches) were constructed for each probe: target alignment. These arrays were then added to the corresponding target array at the start position of the alignment. Repeating this for every probe: target alignment yields the tallied coverage of each nucleotide position by probes. This coverage array was used to compute summary statistics.

Probe synthesis

The starting material for the in-house synthesis of probes is a custom oligo pool (Twist Biosciences, San Francisco, CA). The synthesis scale of the pool ordered will vary based on the total number of unique probes in the pool. For allCARD, a synthesis scale that allowed 32,000–36,000 unique sequences was ordered, while for clinicalCARD, this scale was 16,000–20,000. Each unique oligo in this pool's design contains an individual probe sequence between two other sequences that are the same across all oligos, such that the final length is 120 nt. The first of these flanking sequences is a T7 transcription start site with three extra guanines to reduce transcription efficiency variability (48). The second is a unique primer sequence chosen by the CARPDM program to minimize the probability of interactions with the probe sequences of the pool. This primer must contain an LguI cut site directly proximal to the unique probe sequence, such that it can be cleaved after amplification of the pool and before T7 transcription (Fig. S3).

The last part of CARPDM's pipeline is to convert the designed probe sequences that would typically be ordered as a probeset from a commercial supplier into oligo pool sequences that can be ordered from Twist Biosciences. To do this, CARPDM appends a T7 transcription start site with three extra guanines to one end of every probe sequence. It then creates every possible primer with a terminal LguI cut site of the correct length to yield a final oligo of 120 nt when appended to the opposite side. It compares these putative primers to the concatenated T7/probe sequences using BLASTN and selects the primer with the fewest matches as the second amplification primer, appending the reverse complement to the opposite side of each probe sequence. The resulting sequences were ordered as an oligo-pool from Twist Biosciences (San Francisco, CA), and the amplification primers were ordered from Integrated DNA Technologies (Coralville, IA).

For probe synthesis, sixteen 50 µL PCRs were conducted in parallel, each with 1 ng oligo pool input using 0.5 µL Phusion polymerase with HF buffer (Thermo Fisher Scientific, Waltham, MA), 1 µM of each primer, and 0.2 mM dNTPs (Thermo Fisher Scientific, Waltham, MA). Cycling conditions were initial denaturation at 98°C for 30 s, 12 cycles of 98°C for 10 s, 60°C for 30 s, 72°C for 15 s, and final extension at 72°C for 10 m. Reactions were purified with the QIAQuick Nucleotide Removal Kit (Qiagen, Hilden, Germany), pooling eight reactions per column and eluting each in 30 µL. The elutions were then pooled, and their concentrations were quantified via Qubit 1× dsDNA HS assay (Thermo Fisher Scientific, Waltham, MA).

Four 50 µL restriction endonuclease treatments were then performed in parallel, each with 2 µg PCR input and 2 µL FastDigest LguI (Thermo Fisher Scientific, Waltham, MA). These reactions were then incubated at 37°C for 2 hours, followed by heat inactivation at 65°C for 5 min. Parallel reactions were then pooled over a single Qiagen MinElute PCR

Purification column (Qiagen, Hilden, Germany), eluted in 10 µL, and quantified via Qubit 1× dsDNA HS assay.

Finally, T7 transcription reactions were performed with up to 1 µg of purified LguI digest product, although similar yields were achieved with as little as ~250 ng. For this, HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs, Ipswitch, MA) was used according to the manufacturer's instructions, with one-third of the UTP concentration comprised of Bio-16-UTP (Thermo Fisher Scientific, Waltham, MA). The 20 µL reactions were incubated for 16 h at 37℃, after which 68 µL RNase-free H_2O was added with 10 µL DNase I Buffer and 2 µL DNase I (New England Biolabs, Ipswitch, MA). The resulting mix was incubated for 15 m at 37℃, after which the RNA probes were purified using the Monarch RNA Cleanup kit (50 µg; New England Biolabs, Ipswitch, MA). Finally, concentrations were quantified via Nanodrop (Thermo Fisher Scientific, Waltham, MA), and probes were diluted to 100 ng/µL. One hundred nanograms of each probe set was then analyzed on a 12.5% Urea-PAGE gel stained with SYBR-Gold (Thermo Fisher Scientific, Waltham, MA). A successful reaction was indicated by a smeared band of a slightly higher molecular weight than 80 nt compared to the DNA marker and no remaining band at 120 nt. The probes should appear at a slightly higher molecular weight than 80 nt due to the incorporation of biotin and the extra molecular weight contribution to RNA molecule weight from the 2′ hydroxyl group. The smear indicates the successful incorporation of biotin into the probes depending on uracil content and the inherently stochastic nature of biotinylated vs non-biotinylated uracil insertion. A 120 nt band would indicate the remaining oligo pool if present, which should have been removed by the DNase treatment.

Samples

Two wastewater samples and three soil samples were selected to test the efficacy of the newly constructed probe sets with sequencing. The wastewater samples were 24 h aggregate influent samples from the city of Hamilton (Ontario, Canada) Wastewater Treatment Plant (7 November 2022 and March 2023). DNA was extracted from 50 mL wastewater samples within 24 h of sampling using the DNeasy PowerWater Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols, each with an associated H_2O control. Soil samples were collected from three different environments selected to represent distinct levels of human impact. The first was taken on 10 June 2016, from Holman Island in the Northwestern Territories of Canada, a pristine environment with little human influence. The second was taken on 6 March 2023, from a local wetland in an urban setting in Hamilton (Ontario, Canada), representing an environment with middling human impact. The third was taken on 27 February 2023, from a high-traffic pedestrian area frequented by smokers outside of a Hamilton (Ontario, Canada) hospital, a setting with heavy human influence. All soil samples were stored at −80℃ until processing. DNA was extracted from 250 mg soil samples using the Qiagen DNeasy Powersoil Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols alongside an associated H_2O control.

Sample processing, enrichment, and DNA sequencing

Commercially synthesized probes (CARD v1.0.1 probe set only) and reagents for all enrichments were purchased from Daicel Arbor Biosciences (Ann Arbor, MI). In addition, probes for the allCARD, clinicalCARD, and CARD v1.01 sets were synthesized in-house as outlined above. While allCARD and clinicalCARD probe sets were only synthesized in-house, commercially synthesized CARD v.1.0.1 probes allowed direct comparison with the identical CARD v.1.0.1 probe set synthesized in-house.

For each sample, extracted DNA was quantified via NanoDrop (Thermo Fisher Scientific, Waltham, MA), diluted, and sonicated to an average size of 400 bp using Covaris G-tubes (Woburn, MA). From this, libraries were prepared in quadruplicate from 1 µg DNA using the NEBNext Ultra II ligation kit (New England Biolabs, Ipswich, MA) according to the manufacturer's protocols. These libraries were then pooled before

redistributing to half-size indexing reactions using NEBNext Multiplex Oligos for Illumina (New England Biolabs, Ipswitch, MA). Sample libraries were then enriched in two batches. The first batch performed allCARD and clinicalCARD enrichments, while the second batch performed commercial and in-house CARD v1.0.1 enrichments (Fig. S4). All enrichments were performed according to the manufacturer's protocols (version 5.02) using the Daicel Arbor myBaits v5 kit and reagents in a half-size reaction format with a 24 h hybridization at 62°C, maximum library input, 250 ng of probes, and 14 cycles of post-enrichment reamplification. In-house probes were diluted with RNase-free water to the same concentration as Daicel Arbor's (100 ng/µL) before use in the protocol. Each enriched library was compared to the same sample without enrichment.

Before sequencing, libraries were quantified in triplicate using NEB Luna Universal Probe qPCR Master Mix (New England Biolabs, Ipswitch, MA), Illumina PhiX standard (San Diego, California), and the following primers, all ordered from Integrated DNA Technologies (Coralville, IA): P5: AATGATACGGCGACCACCGA, P7: CAAGCAGAAGACGGCATACGA, and probe: /56-FAM/CCCTACACG/ZEN/ACGCTCTTCCGATCT/3IABkFQ/. Cycling conditions were initial denaturation at 95°C for 3 m, followed by 40 cycles of denaturation at 95°C for 15 s and annealing/extension at 60°C for 1 m. The resulting concentrations were used to make individual pools for each enrichment set above, alongside one replicate of the shotgun (without enrichment) libraries (Fig. S4). Paired-end reads of 2 × 150 bp were obtained from each pool using an Illumina NextSeq 2000 (San Diego, California), with samples having an average depth of 4.47 M clusters sequenced (minimum of 3.26 M clusters and a maximum of 6.71 M clusters).

Analysis and visualization

All analyses were performed with custom Python scripts and visualized with the ggplot2 (49) package in R. For rarefaction analysis, libraries were subsampled every 100,000 paired-end reads up to 3 M using seqtk v1.3. Fastp v0.23.2 28 performed initial read trimming, quality control, and deduplication for these subsamples without merging. CARD's RGI bwt v6.0.2 tool with KMA v1.4.9 (50) then mapped reads to reference sequences in CARD v3.2.5. ARGs were classified as present if they had more than 100 reads mapping (regardless of the breadth of coverage of the reference sequence). Regression lines were determined without extrapolation using the ggplot local estimated scatterplot smoothing function with geom_smooth. Since one cannot extrapolate a locally estimated function, we used a standard log-linear model when extrapolating the predicted number of ARGs detected at a sequencing depth of 10 M. GNU parallel v20161222 (51), the Python pandas v1.5.3 library (52), and BioPython v1.78 (53) were used heavily for these analyses.

For read distribution analysis, custom Python scripts used the CARD RGI bwt output to count the number of reads that mapped to each ARG in soil and wastewater samples, respectively. The top 20 most prevalent ARGs (i.e., ARGs with the highest number of reads mapping) in each sample source (soil and wastewater) were kept for the figure, while all others were collapsed to the "other" category. This cut-off was chosen to keep figure legends readable while providing discriminatory power between the performance of different probe sets. Mean counts of the two replicates are used to determine the number plotted. The distribution of percent identity between read and reference for each sample was determined by a custom Python script that parsed the CIGAR strings in the BAM file that accompanies the RGI bwt output, where the percent identity was calculated as the number of nucleotide matches/151 × 100. Replicates were pooled for this analysis.

In a clinicalCARD vs allCARD overlap analysis, a custom Python script determined clinically relevant ARGs detected by allCARD or clinicalCARD with ≥100 mapped reads in both replicates at each subsampling depth. A similar approach was used to determine the overlap between the in-house vs commercially synthesized CARD v1.01 probe sets. However, to compare these sets, only ARGs included in the initial design (i.e., in CARD v1.0.1) were considered. Coverage analysis only considered clinically relevant ARGs

detected by both clinicalCARD and allCARD with at least one read at a subsampled sequencing depth of 3 M paired-end reads in both replicates. Python scripts determined each ARG's coverage using the CARD RGI bwt output from the rarefaction analysis. Once again, for a similar analysis comparing the commercial and in-house synthesis, only ARGs from CARD v1.0.1 were considered.

For a wastewater read correlation analysis, only clinically relevant (i.e., in clinicalCARD) ARGs detected with at least one read in both shotgun and enriched in each sample were considered. All ARGs were included from soil samples, while clinicalCARD analysis was omitted due to the sparsity of detected ARGs. R^2 values were calculated using scikit-learn v1.2.2 (54). The cmlA1 read number correction outlined below was accomplished by summing the number of reads attributed to all closely related *cmlA* variants (i.e., cmlA1, 4, 5, 6, 8, and 9) in each relevant treatment and manually adding the corresponding values to the plot.

RESULTS

Probe design and synthesis

After design and filtering, allCARD contained 34,915 unique probe sequences covering 4,661 ARGs. Alternatively, clinicalCARD contained 15,393 unique probes, covering 323 ARGs (Fig. S5). No probe set had zero coverage of an ARG against which it was designed. When analyzing probes against all ARGs curated as CARD protein homolog models (Fig. 2a), the median value of all metrics other than the proportion covered per ARG was higher in the clinicalCARD probe set. The reason for clinicalCARD's bimodal distribution compared to allCARD is that it was only designed against the clinically relevant subset. Therefore, the clinically irrelevant genes (e.g., *tolC* and *H-NS*) had no probes aligning since they were not included in the initial design. However, despite only being designed against 323 ARGs, over 50% of CARD ARGs had coverage of >75% by clinicalCARD. When analyzing both probe sets against the clinically relevant ARGs in clinicalCARD (Fig. 2b), median values in clinicalCARD were higher than allCARD in every metric.
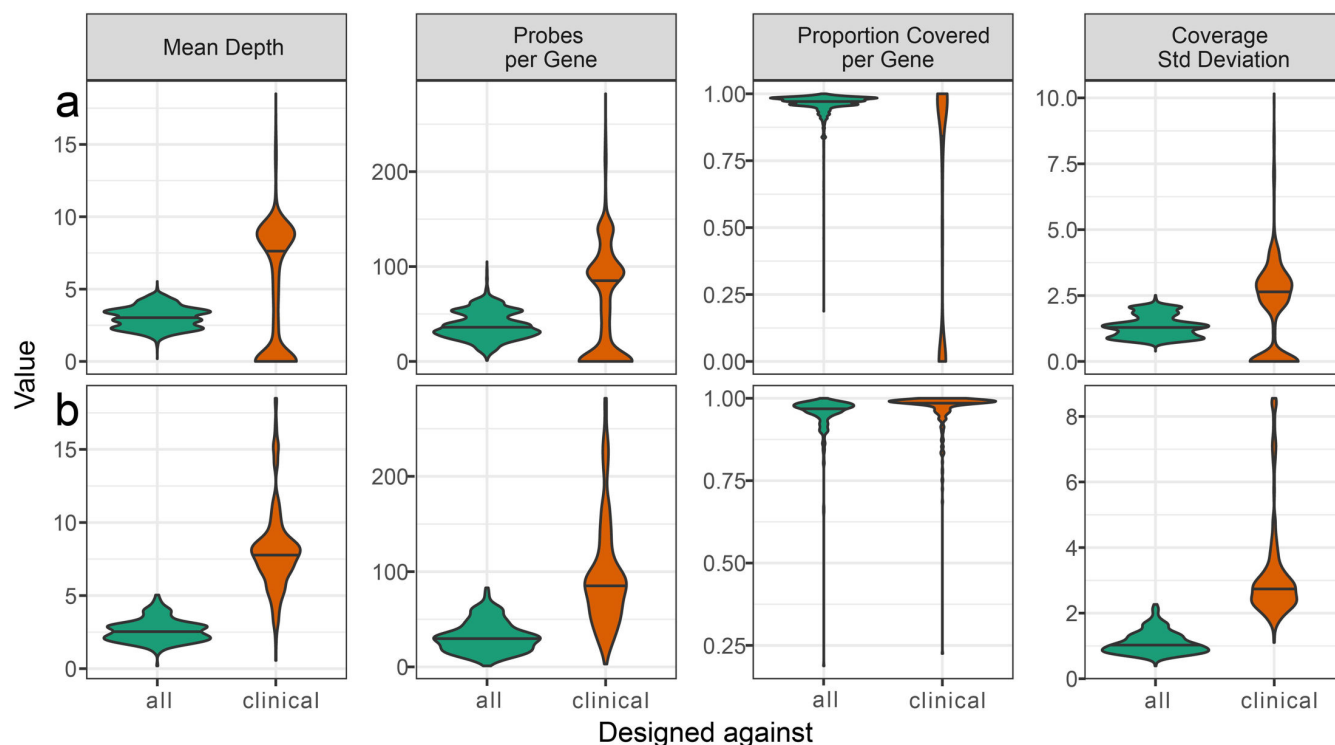

FIG 2 *In silico* probe analysis using BLAST to align all probes in allCARD and clinicalCARD against (a) all genes in CARD and (b) all genes deemed to be clinically relevant.

The increased *in silico* performance of clinicalCARD is a consequence of the smaller initial set of reference sequences. As such, it did not have to remove as many probes during the redundancy filter (Fig. S5a) and ended up with more redundancy and superior coverage of the ARGs against which it was designed. Probes in clinicalCARD had a lower median GC content and melting temperature. A higher proportion of probes also target only a single ARG in clinicalCARD relative to allCARD (Fig. S5b). After synthesis, a urea-PAGE gel shows a smear above 80 nt due to the stochastic incorporation of biotin into the probe (Fig. S6). A comparison of all pools against the commercially synthesized version indicates a similar size distribution, with a slight bias toward lower molecular weights in the in-house synthesized probe set.

Probe testing

Five samples in five different treatments were employed to determine the efficacy of the probes when enriching for targets. No blanks had sufficient sequencing depth to be analyzed at even the lowest subsampling depth, indicating negligible contamination. First, the non-inferiority of the in-house synthesized probe set relative to the commercial option was established. After subsampling every 100 k paired-end reads up to 3 M with analysis by the CARD RGI bwt tool (55), the number of ARGs with >100 mapped reads was determined for each depth in each sample (Fig. 3). This analysis illustrated that the in-house synthesized probes detected more ARGs at the same sequencing depth in every sample. This effect was more pronounced in the soil samples, where the in-house synthesized probe set detected as many as double the number of ARGs detected by
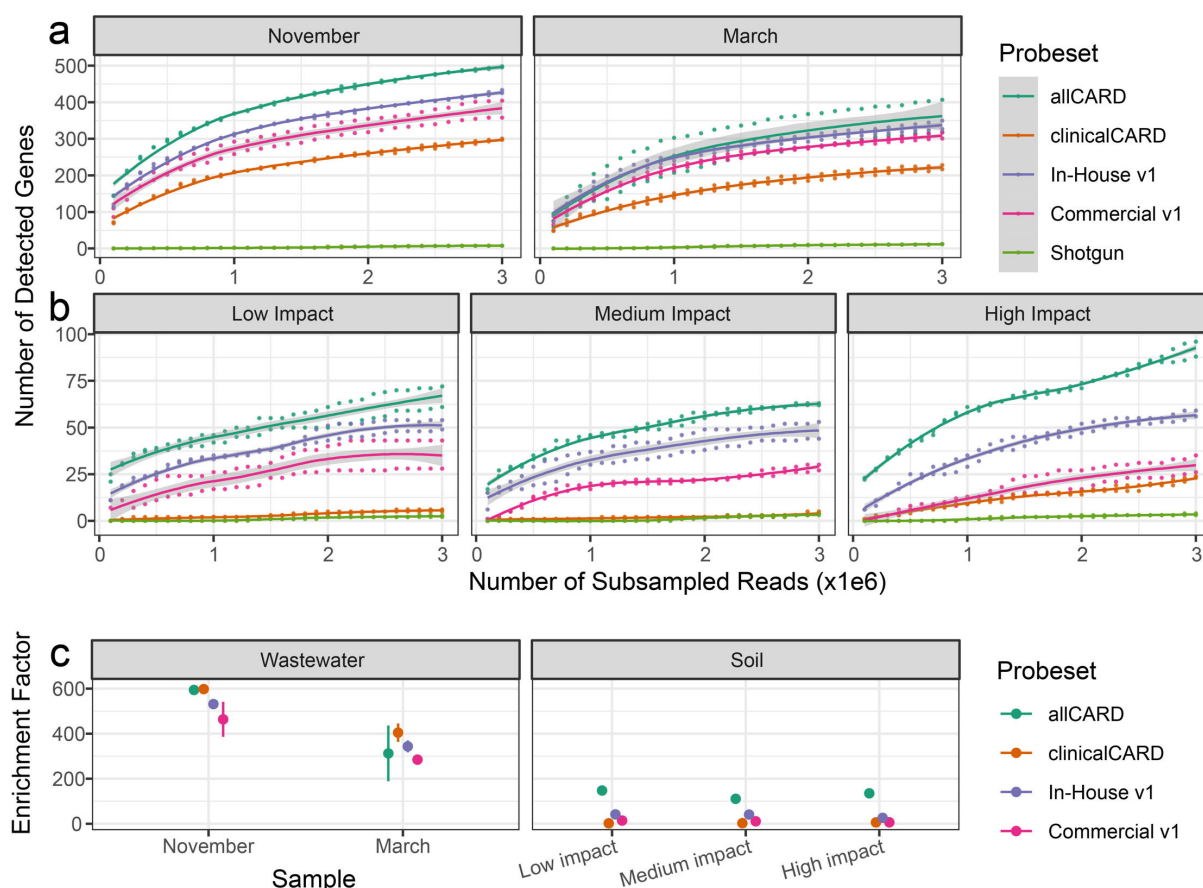

FIG 3 Enrichment efficiency comparisons between different treatments. Rarefaction analysis of the number of detected genes at subsampling depths of every 100 k reads up to 3 M in (a) wastewater samples and (b) soil samples. Shaded areas represent the 95% CI. (c) Average enrichment factor ±1 SD of different probe sets in different samples. The enrichment factor is defined here as the number of CARD-mapped reads after enrichment with the relevant probe set divided by the number of CARD-mapped reads in shotgun sequencing data.

the commercially synthesized probe set, despite both having identical probe sequences based on CARD v1.01. Yet, our in-house synthesized probe set had an almost identical distribution of the top 20 detected ARGs after enrichment compared to the commercially synthesized set (Fig. 4), indicating consistent enrichment efficiency between ARGs in the commercial and in-house synthesized sets. These results were further complemented by analyzing the overlap between ARGs detected by each set at each subsampling depth and the associated coverage distribution for all detected ARGs (Fig. S7 and S8). In these analyses, enrichment with our in-house synthesized probes detected ARGs with less sequencing effort than commercially synthesized probes and had greater coverage of detected ARGs.

Enrichment with all probe sets detected vastly more ARGs in wastewater samples than by sequencing without enrichment. allCARD detected the most, with up to 498 different ARGs, and clinicalCARD detected the least, with up to 300 ARGs. However, clinicalCARD efficiently retained the highest abundance of ARGs, evidenced by the similar distributions between allCARD and clinicalCARD in Fig. 4a. allCARD detected by far the greatest number of ARGs in soil samples, up to 96 in the high-impact hospital grounds. clinicalCARD did not detect a substantial number of ARGs in any soil sample, except for that taken from the high-impact hospital grounds, where it detected 24.

When subsampled to the same depth, the enrichment factor (i.e., the number of enriched reads mapped to CARD divided by the number of shotgun reads mapped to CARD) of different probe sets shows a more consistent value within each sample
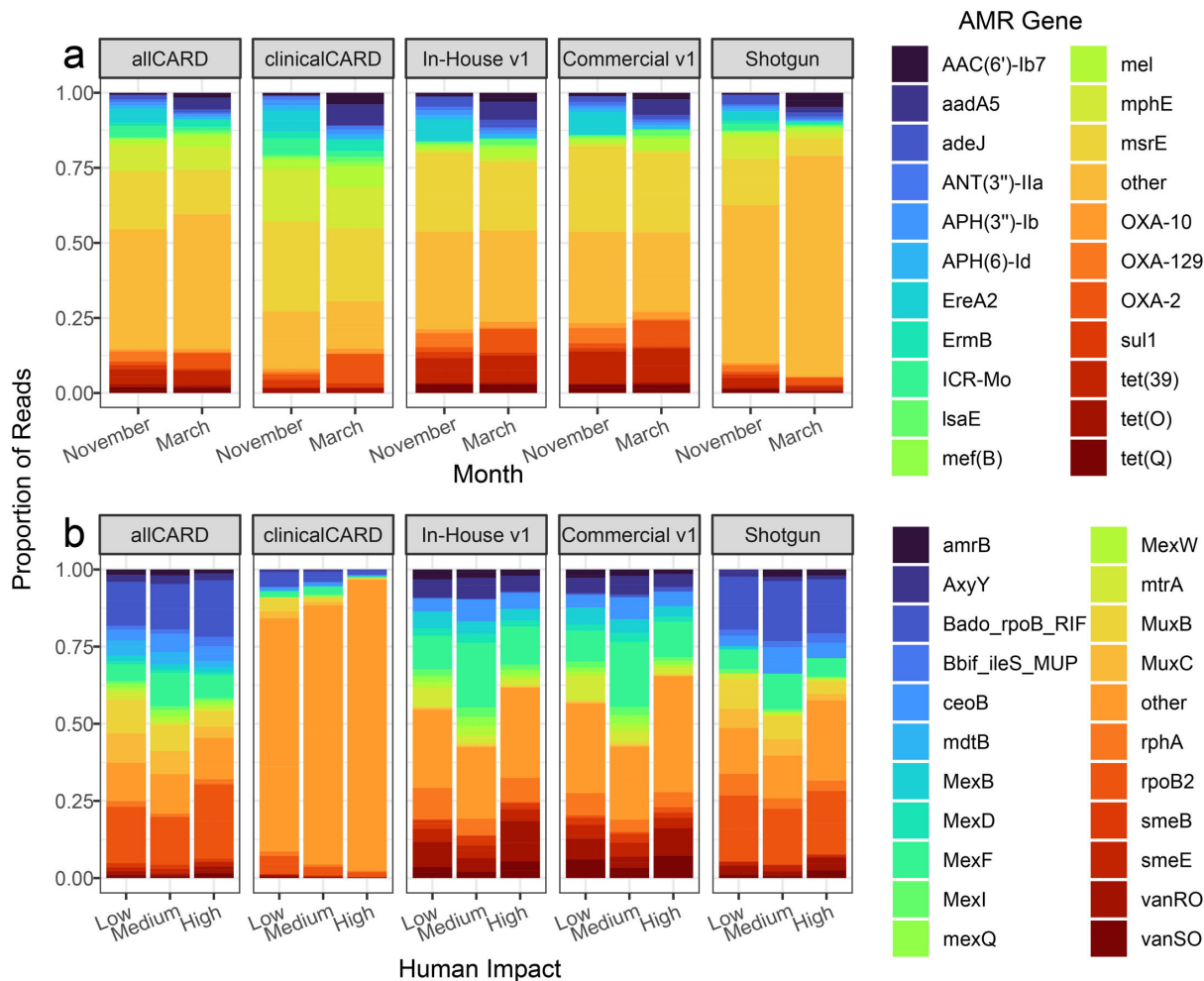

FIG 4 Top 20 AMR genes according to overall count and their associated read proportions in each sample: (a) wastewater samples and (b) soil samples. Very long gene names were replaced with their CARD short name.

than expected, given the high performance of allCARD at detecting the most ARGs (Fig. 3c). This is especially evident in wastewater, where clinicalCARD is on par with, if not outperforming, allCARD in terms of enrichment factor despite detecting far fewer ARGs in the rarefaction curve. There was a 400–600-fold enrichment in the wastewater samples in the November sample, depending on the probe set, which decreased to 200–400-fold for the March sample. Soil samples had a markedly lower enrichment factor than wastewater, hovering between 0- and 200-fold enrichment depending on the probe set. Yet, enrichment was consistent between samples for different probe sets in soil. Based on the rarefaction curves (Fig. 3), ARG detection in most samples begins to plateau by a sequencing depth of 3 M paired-end reads. Extrapolation of these rarefaction curves to a sequencing depth of 10 M paired-end reads indicates that at a subsampling depth of 3 M, we have captured 70%–80% of the diversity of ARGs that may be detected at 10 M (Fig. S9).

There were differences when comparing the top 20 most prevalent ARGs in soil and wastewater (Fig. 4). First, all top 20 ARGs in wastewater data, save for *adeJ* and *tetQ*, were in our set of clinically relevant ARGs. However, in soil samples, not a single top 20 ARG was included in the clinically relevant set, which explains the lack of enrichment for any top 20 ARGs in soil by the clinicalCARD probe set. Most dominant ARGs in the soil samples were found across various species and associated with efflux pumps (e.g., *mex* genes) or transcriptional regulators (e.g., *mtrA* and *vanR/S*). Moreover, the percent identity of the ARG-associated reads relative to the CARD reference sequences in the soil samples was lower than those from wastewater samples (Fig. 5). Before enrichment, the wastewater sample had a mixture of high- (>90%) and mid-identity (60%–90%) reads relative to their references in CARD, but the soil samples had exclusively mid- to low-identity (<60%) reads. After enrichment, wastewater samples had almost exclusively high-identity reads in both the clinicalCARD and allCARD enrichments. In soil samples, enrichment with allCARD was selected heavily for mid-identity reads, but clinicalCARD was preferentially selected for high-identity reads, especially in the sample from the high-impact hospital grounds.
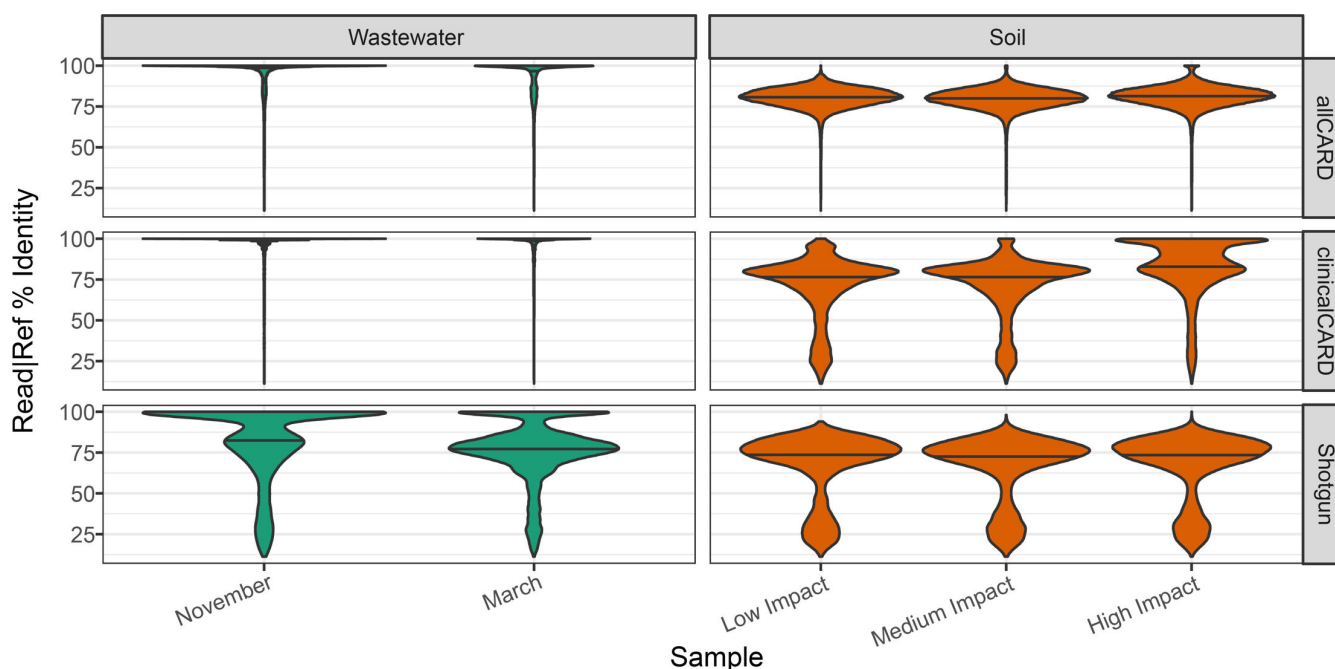

FIG 5 Distribution of the number of similarities between read and reference, based on CIGAR strings from the BAM file output by the CARD RGI *bwt* module. The internal line of each violin plot indicates the median of the distribution.

allCARD vs clinicalCARD

The wastewater samples were used to examine clinicalCARD's efficacy relative to allCARD at enriching the clinically relevant ARGs it was designed against, as clinically relevant ARGs had the highest abundance in the wastewater samples. Overall, clinicalCARD detected an average of 16% more ARGs than allCARD in the November wastewater sample and 35% more in the March sample at each subsampling depth, with very rarely an ARG detected uniquely by allCARD (Fig. 6a).

To assess the breadth of coverage of individual ARGs after enrichment with allCARD or clinicalCARD, clinically relevant ARGs detected by both probe sets in both replicates at a subsampling depth of 3 M paired-end reads were analyzed. The distribution of coverage of these ARGs at each subsampling depth for allCARD and clinicalCARD was plotted (Fig. 6b). clinicalCARD delivered 100% coverage of at least 50% of the ARGs detected in the November sample at a subsampling depth of 1 M reads, while allCARD took 1.5 M reads to do the same. In the March sample, clinicalCARD took only 700 k reads to reach this level of detection, while allCARD required 2.1 M reads. Based on this analysis, clinicalCARD delivers better coverage of more clinically relevant ARGs at a lower sequencing depth than allCARD.

Finally, a commonly perceived limitation of enrichment is that hybridization can be sequence dependant, which may introduce biases in the final library, thereby eliminating the ability to perform relative quantification of ARGs in a sample. To investigate the relationship between ARG abundance in enriched vs unenriched data, the clinically relevant ARGs present with at least one read in both replicates of shotgun and enriched data at a subsampling depth of 3 M paired-end reads were analyzed (Fig. 7). In the wastewater samples, due to the high identity between reads, the R^2 value reached as
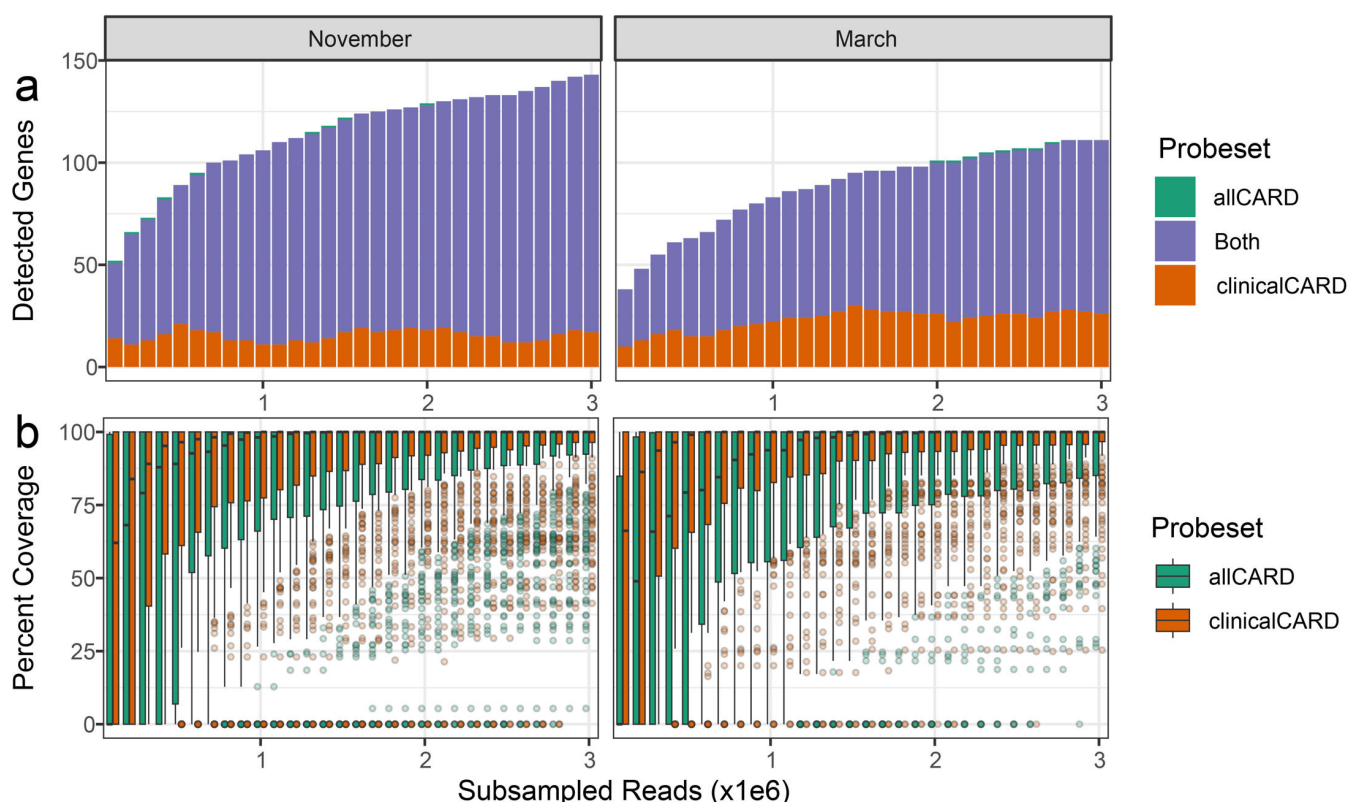

FIG 6 Comparison of enrichment efficiency of clinically relevant ARGs between allCARD and clinicalCARD probe sets. (a) Overlap analysis showing ARGs detected by allCARD and clinicalCARD, for both with at least 100 reads at different subsampling depths. (b) Coverage analysis showing the distribution of coverage for ARGs detected by both probe sets with at least one read at a depth of 3 M. In the November sample, there were 226 such ARGS. In the March sample, there were 176.
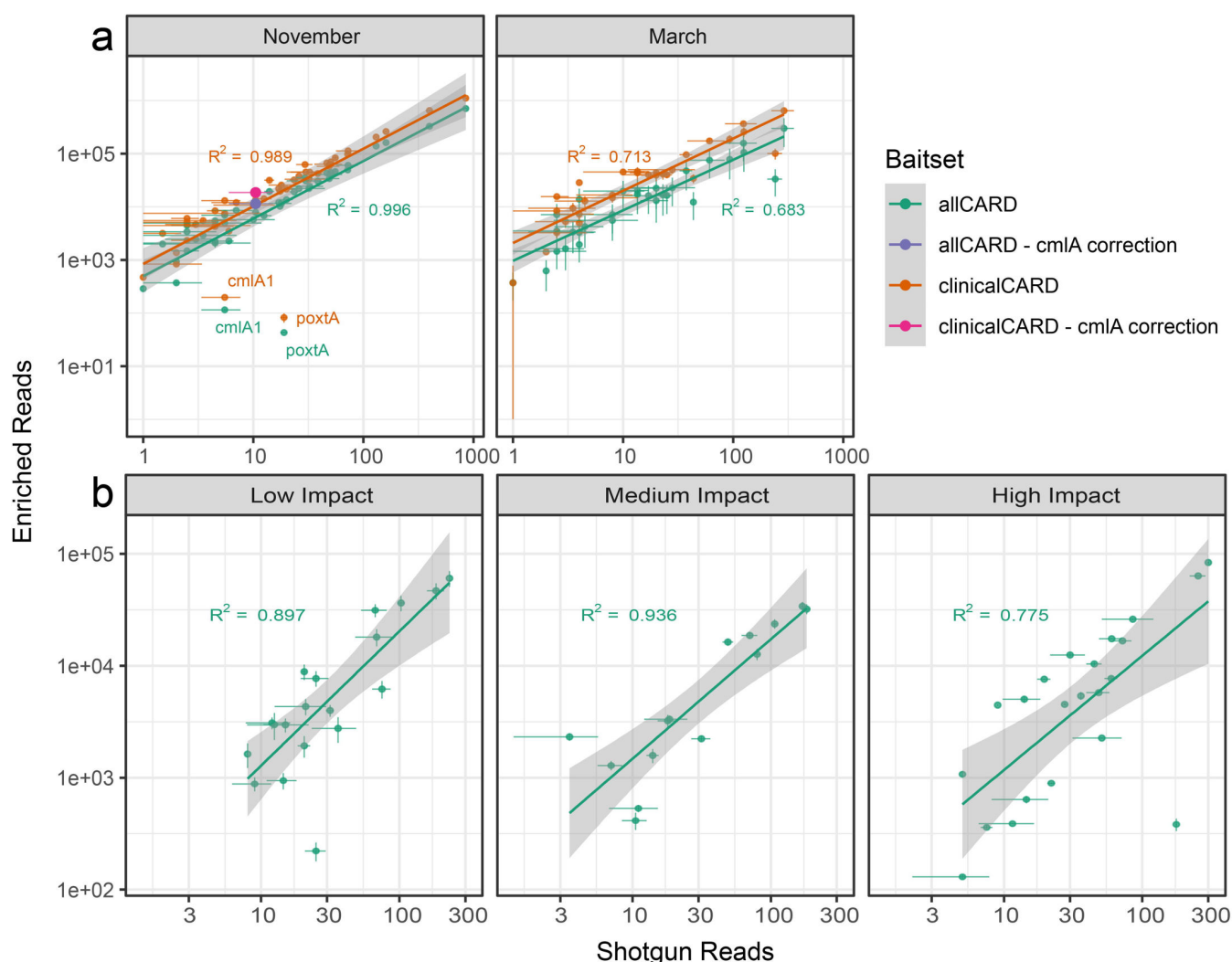
FIG 7 Correlation of the number of enriched and shotgun reads for each ARG with associated R^2 values. Shaded areas represent the 95% CI. (a) Wastewater analysis: 44 and 31 ARGs fit the inclusion criteria for the November and March samples, respectively. The purple and pink dots show cmlA1's position for allCARD and clinicalCARD enrichment, respectively, if all reads from all extremely closely related cmlA variants were considered to be from cmlA1. (b) Soil analysis: clinicalCARD was not analyzed due to a lack of reads mapping to clinically relevant ARGs. In the low human impact sample, 19, 14, and 22 ARGS fit the inclusion criteria for the low human impact, medium impact, and high impact samples, respectively.

high as 0.996 in the November sample with the allCARD probe set, while it was slightly lower after enrichment with clinicalCARD. In the March wastewater sample, R^2 values were considerably lower, most likely due to the inconsistent replicates. Enrichment was generally less efficient in the soil samples due to the lower identity between DNA and probes; however, the R^2 values remained high, reaching 0.897 in the low-impact sample, 0.936 in the medium-impact, and 0.775 in the high-impact sample from the hospital grounds.

In Fig. 7, ARGs were labeled on the plot if the enrichment factor was significantly lower ($P < 0.05$) than the average enrichment factor within a sample and probe set. Two ARGs in the November wastewater sample were obvious outliers. The first, *cmlA1*, has several close homologs with >99% nucleotide identity to which thousands of reads were assigned. This occasional occurrence of reads mapping among highly similar alleles is known as the allele network problem (50, 56). When mapping to highly redundant databases, even with new tools such as KMA (50) as used by RGI bwt, reads can be misassigned to another closely related allele or ARG. To illustrate this, we superimposed

corrected points onto the plot, i.e., showing where cmlA1 would reside on the plot if all the reads attributed to its variants were attributed to it instead. As expected, this correction brings it into far better agreement with the observed trend. The reads of the second outlier ARG, *paxtA*, appear to be from a distant homolog of the reference in CARD. Due to this sequence variance, it has a lower enrichment efficiency.

DISCUSSION

Our work has resulted in the design of two probe sets to enrich ARGs in metagenomic sequencing libraries. Several improvements have been made relative to the first iteration of this probe set against the first version of CARD. First, a denser initial tiling and exclusion of perfect matches during the nt BLAST filter maximized coverage against all ARGs. Second, a stringent redundancy filter minimized unneeded probes. This enables allCARD to enrich all ARGs included in CARD more evenly with fewer probes. Alternatively, clinicalCARD has a much higher redundancy since it was designed against a much smaller set of clinically relevant ARGs. Our in-house synthesis protocol yields a similar smear pattern to the commercially synthesized version when run on a urea-PAGE gel, indicating similar physical properties. There was a slight bias to lower molecular weights in the in-house synthesized probe sets, which may indicate more incorporation of biotin into the commercial probes.

When comparing *in silico* statistics of allCARD vs clinicalCARD, it is evident that clinicalCARD outperforms allCARD when enriching for the ARGs it was designed against. clinicalCARD is superior at targeting the ARGs against which it was designed due to its decreased need for redundancy filtering during probe design. Since it covers a smaller number of ARGs, fewer probes are made during the initial tiling step with BaitsTools, and fewer cycles of redundancy filtering are needed to satisfy the total probe number cut-off (Fig. 2a). This translates to more probes per ARG and, therefore, better coverage of the ARGs against which it was designed. However, clinicalCARD also has coverage against a much larger portion of CARD than only those ARGs (i.e., median coverage of clinicalCARD probes against allCARD ARGs is >75%). This is because of the high degree of conservation of ARG nucleotide sequences and the resulting redundancy in CARD, where one ARG may have several closely related variants (35). Overall, clinical-CARD has higher median scores in all tested metrics against its design set of ARGs. Based on this, clinicalCARD should be more effective when enriching clinically relevant ARGs. Additionally, since it is a smaller set of probes, it will be less expensive, even with in-house synthesis. However, it lacks coverage of many ARGs in allCARD. As some questions require comprehensive coverage of all ARGs, allCARD is the better choice in these situations.

To test our new probe sets, five samples were used for testing: two wastewater samples spanning the beginning and end of winter and three soil samples from differently human-impacted sites. In all samples, enrichment drastically improved the detection of ARGs relative to shotgun sequencing, with an enrichment factor of up to 598-fold. Our in-house synthesis consistently detected more ARGs at a lower sequencing depth than the commercially synthesized probes. This may be due to the evenness of coverage in the Twist Biosciences oligo-pool, which may provide a superior template for transcription and, thus, a superior capture reagent.

In wastewater samples, allCARD detected the most ARGs by a considerable margin. clinicalCARD detected relatively few ARGs, although those it did detect were among the most frequent ARGs in the sequencing data. allCARD once again detected the most ARGs in the soil samples, but clinicalCARD missed many ARGs in the soil except in the sample from the high-impact hospital grounds. This is likely due to the lack of clinically relevant ARGs in soil samples from environments with less human impact. We stress, however, that this experiment, as designed, does not indicate causality from being near a hospital. It merely supports the notion that samples from environments with high human influence carry more ARGs, a known phenomenon (57, 58).

Compared to the large fluctuation in the number of ARGs detected between probe sets, the enrichment factor among probe sets was more consistent, especially in the wastewater samples. This suggests that with a certain input of probes, one can expect a given level of enrichment so long as the target is sufficiently abundant. Enrichment suffers when this is not the case, such as in the non-allCARD probe sets for the soil samples. When inspecting the number of identities between reads and references, there is a pronounced difference between wastewater and soil samples. Without enrichment, the distribution of the number of identities in unenriched reads of wastewater samples indicates the presence of both distant and close homologs to the references in CARD. After enrichment, only close homologs are present, suggesting that they are preferentially enriched. However, distant homologs are enriched in their absence, such as in the low- and medium-impact soil samples (Fig. 5b). However, in the sample from the high-impact hospital grounds, there was a bias toward close homologs, especially in the sample enriched by clinicalCARD. Overall, this shows a pattern where clinically relevant ARGs have a higher identity to their references in CARD and thus are more efficiently enriched, particularly given that CARD and other ARG databases are biased toward clinical isolates (35). The distant homologs in soil samples are mainly associated with efflux and regulation, and we cannot know that these distant homologs confer the same degree and type of AMR as their reference in CARD. For this, experimental data are required. However, we can conclude that environmental samples with low human impact are less likely to contain exact matches to reference ARGs in CARD. For this reason, these samples are less likely to perform as well when being enriched.

There was a striking linearity when considering the correlation between the number of ARG-associated reads in enriched vs unenriched samples. This was unexpected due to the assumed effect that sequence differences, specifically GC content, would have on hybridization efficiency (59, 60). Relative quantification of ARGs within a sample and comparison among samples may be possible, i.e., a large difference in read abundance in enriched data indicates a proportionally large difference in shotgun data, although smaller differences may be challenging to detect reliably. However, our investigation contains few samples, and quantifying this relationship was not our primary aim.

There are limitations to enrichment as a method for resistome profiling. Enrichment cannot detect entirely novel resistance genes that are not in CARD. As such, we designed CARPDM to update the probe set for each released version of CARD. Combined with reduced up-front costs via our novel synthesis method, researchers can update their probe set on demand. Moreover, when the genes that get enriched are distant homologs to those in CARD, we cannot be sure if they are actual resistance genes or if they would generate clinical levels of resistance. More work is required to validate these genes upon detection. Finally, this protocol relies on proprietary hybridization reagents and buffers from commercial suppliers.

Conclusions

The increasing global burden of AMR requires cost-effective and scalable solutions. Overall, our in-house synthesized probes detect more ARGs with less sequencing than a commercial option in the samples tested. allCARD robustly enriches the vast array of ARGs against which it was designed, as well as their distant homologs. clinicalCARD even more robustly enriches the smaller set of clinically relevant ARGs against which it was designed. This work shows that targeted enrichment is a valuable companion to DNA sequencing when detecting ARGs. Future work will involve further updating and refining the set of clinically relevant ARGs and continually updating the probe set with every new CARD release. Additionally, while we have released a preliminary protocol for the in-house synthesis of any probe set, there remains room for this protocol to be optimized and made even less reliant on commercial reagents. Finally, to make this technology more accessible, future work should also investigate its efficacy on alternative DNA sequencing platforms, such as Oxford Nanopore's MinION. Yet, overall, this work has shown the power of enrichment to decrease the cost and increase the

impact of large-scale monitoring of ARGs using DNA sequencing. Alongside CARD, this technology can help researchers investigate ARG prevalence and transmission patterns among different populations and environments.

AUTHOR AFFILIATIONS

[1]M.G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada
[2]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada
[3]McMaster Ancient DNA Center, McMaster University, Hamilton, Ontario, Canada
[4]Department of Anthropology, McMaster University, Hamilton, Ontario, Canada
[5]Humans and the Microbiome Program, Canadian Institute for Advanced Research, Toronto, Ontario, Canada

AUTHOR ORCIDs

Gerard D. Wright ⓘ http://orcid.org/0000-0002-9129-7131
Andrew G. McArthur ⓘ http://orcid.org/0000-0002-1142-3063

AUTHOR CONTRIBUTIONS

Dirk Hackenberger, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing | Hamna Imtiaz, Data curation, Investigation, Methodology, Software, Writing – review and editing | Amogelang R. Raphenya, Methodology, Software, Writing – review and editing | Brian P. Alcock, Data curation, Methodology, Software, Writing – review and editing | Hendrik N. Poinar, Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review and editing | Gerard D. Wright, Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review and editing | Andrew G. McArthur, Conceptualization, Funding acquisition, Project administration, Resources, Software, Supervision, Writing – review and editing

DATA AVAILABILITY

Detailed laboratory protocols and FASTA sequence files for probe synthesis and library enrichment are available at the Comprehensive Antibiotic Resistance Database

website: http://card.mcmaster.ca. Version controlled copies of the CARPDM bioinformatics workflow are available in GitHub repository https://github.com/arpcard/carpdm. All metagenomic sequencing results, enriched or shotgun, were deposited in NCBI under BioProject PRJNA1093140.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental figures (AEM01876-24-s0001.pdf). Figures S1 to S9.
Supplemental text (AEM01876-24-s0002.txt). Genes and their sequences curated for clinicalCARD probe design.

REFERENCES

1. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, Gray A, Han C, Bisignano C, Rao P, Wool E, et al. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet 399:629–655. https://doi.org/10.1016/S0140-6736(21)02724-0
2. O'Neill J. 2014. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. The Review on Antimicrobial Resistance, London.
3. Waglechner N, Culp EJ, Wright GD. 2021. Ancient antibiotics, ancient resistance. EcoSal Plus 9:eESP-0027-2020. https://doi.org/10.1128/ecosalplus.ESP-0027-2020
4. Waglechner N, McArthur AG, Wright GD. 2019. Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. Nat Microbiol 4:1862–1871. https://doi.org/10.1038/s41564-019-0531-5
5. Hall BG, Barlow M. 2004. Evolution of the serine β-lactamases: past, present and future. Drug Resist Updat 7:111–123. https://doi.org/10.1016/j.drup.2004.02.003
6. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. Nat Genet 28:281–285. https://doi.org/10.1038/90129
7. Barlow M, Hall BG. 2002. Phylogenetic analysis shows that the OXA β-lactamase genes have been on plasmids for millions of years. J Mol Evol 55:314–321. https://doi.org/10.1007/s00239-002-2328-y
8. Lee JH, Takahashi M, Jeon JH, Kang L-W, Seki M, Park KS, Hong M-K, Park YS, Kim TY, Karim AM, Lee J-H, Nashimoto M, Lee SH. 2019. Dual activity of PNGM-1 pinpoints the evolutionary origin of subclass B3 metallo-β-lactamases: a molecular and evolutionary study. Emerg Microbes Infect 8:1688–1700. https://doi.org/10.1080/22221751.2019.1692638
9. Larsson DGJ, Flach CF. 2022. Antibiotic resistance in the environment. Nat Rev Microbiol 20:257–269. https://doi.org/10.1038/s41579-021-00649-x
10. Poirel L, Rodriguez-Martinez JM, Mammeri H, Liard A, Nordmann P. 2005. Origin of plasmid-mediated quinolone resistance determinant QnrA. Antimicrob Agents Chemother 49:3523–3525. https://doi.org/10.1128/AAC.49.8.3523-3525.2005
11. Cantón R, González-Alba JM, Galán JC. 2012. CTX-M enzymes: origin and diffusion. Front Microbiol 3:110. https://doi.org/10.3389/fmicb.2012.00110
12. Ebmeyer S, Kristiansson E, Larsson DGJ. 2021. A framework for identifying the recent origins of mobile antibiotic resistance genes. Commun Biol 4:8. https://doi.org/10.1038/s42003-020-01545-5
13. Wright GD. 2007. The antibiotic resistome: the nexus of chemical and genetic diversity. Nat Rev Microbiol 5:175–186. https://doi.org/10.1038/nrmicro1614
14. Hendriksen RS, Munk P, Njage P, van Bunnik B, McNally L, Lukjancenko O, Röder T, Nieuwenhuijse D, Pedersen SK, Kjeldgaard J, et al. 2019. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun 10:1124. https://doi.org/10.1038/s41467-019-08853-3
15. Munk P, Brinch C, Møller FD, Petersen TN, Hendriksen RS, Seyfarth AM, Kjeldgaard JS, Svendsen CA, van Bunnik B, Berglund F, Global Sewage Surveillance Consortium, Larsson DGJ, Koopmans M, Woolhouse M, Aarestrup FM. 2022. Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. Nat Commun 13:7251. https://doi.org/10.1038/s41467-022-34312-7
16. Chau KK, Barker L, Budgell EP, Vihta KD, Sims N, Kasprzyk-Hordern B, Harriss E, Crook DW, Read DS, Walker AS, Stoesser N. 2022. Systematic review of wastewater surveillance of antimicrobial resistance in human populations. Environ Int 162:107171. https://doi.org/10.1016/j.envint.2022.107171
17. Harrington A, Vo V, Papp K, Tillett RL, Chang C-L, Baker H, Shen S, Amei A, Lockett C, Gerrity D, Oh EC. 2022. Urban monitoring of antimicrobial resistance during a COVID-19 surge through wastewater surveillance. Sci Total Environ 853:158577. https://doi.org/10.1016/j.scitotenv.2022.158577
18. Majeed HJ, Riquelme MV, Davis BC, Gupta S, Angeles L, Aga DS, Garner E, Pruden A, Vikesland PJ. 2021. Evaluation of metagenomic-enabled antibiotic resistance surveillance at a conventional wastewater treatment plant. Front Microbiol 12:657954. https://doi.org/10.3389/fmicb.2021.657954
19. Yin X, Deng Y, Ma L, Wang Y, Chan LYL, Zhang T. 2019. Exploration of the antibiotic resistome in a wastewater treatment plant by a nine-year longitudinal metagenomic study. Environ Int 133:105270. https://doi.org/10.1016/j.envint.2019.105270
20. Szczepanowski R, Linke B, Krahn I, Gartemann KH, Gützkow T, Eichler W, Pühler A, Schlüter A. 2009. Detection of 140 clinically relevant antibiotic-resistance genes in the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to selected antibiotics. Microbiology (Reading) 155:2306–2319. https://doi.org/10.1099/mic.0.028233-0
21. Osbiston K, Oxbrough A, Fernández-Martínez LT. 2021. Antibiotic resistance levels in soils from urban and rural land uses in Great Britain. Access Microbiol 3:acmi000181. https://doi.org/10.1099/acmi.0.000181
22. Li HZ, Yang K, Liao H, Lassen SB, Su JQ, Zhang X, Cui L, Zhu YG. 2022. Active antibiotic resistome in soils unraveled by single-cell isotope probing and targeted metagenomics. Proc Natl Acad Sci U S A 119:e2201473119. https://doi.org/10.1073/pnas.2201473119
23. Mills M, Lee S, Mollenkopf D, Wittum T, Sullivan SMP, Lee J. 2022. Comparison of environmental microbiomes in an antibiotic resistance-polluted urban river highlights periphyton and fish gut communities as reservoirs of concern. Sci Total Environ 851:158042. https://doi.org/10.1016/j.scitotenv.2022.158042
24. Goodarzi Z, Asad S, Mehrshad M. 2022. Genome-resolved insight into the reservoir of antibiotic resistance genes in aquatic microbial community. Sci Rep 12:21047. https://doi.org/10.1038/s41598-022-25026-3
25. Qian X, Gunturu S, Guo J, Chai B, Cole JR, Gu J, Tiedje JM. 2021. Metagenomic analysis reveals the shared and distinct features of the soil resistome across tundra, temperate prairie, and tropical ecosystems. Microbiome 9:108. https://doi.org/10.1186/s40168-021-01047-4
26. Hu Y, Yang X, Li J, Lv N, Liu F, Wu J, Lin IYC, Wu N, Weimer BC, Gao GF, Liu Y, Zhu B. 2016. The bacterial mobile resistome transfer network connecting the animal and human microbiomes. Appl Environ Microbiol 82:6672–6681. https://doi.org/10.1128/AEM.01802-16
27. Kim DW, Cha CJ. 2021. Antibiotic resistome from the One-Health perspective: understanding and controlling antimicrobial resistance

transmission. Exp Mol Med 53:301–309. https://doi.org/10.1038/s12276-021-00569-z

28. Rimal R, Shrestha D, Pyakurel S, Poudel R, Shrestha P, Rai KR, Ghimire GR, Rai G, Rai SK. 2022. Diagnostic performance of GeneXpert MTB/RIF in detecting MTB in smear-negative presumptive TB patients. BMC Infect Dis 22:321. https://doi.org/10.1186/s12879-022-07287-5

29. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, Wlodarski MA, Edalatmand A, Petkau A, Syed SA, Tsang KK, et al. 2023. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res 51:D690–D699. https://doi.org/10.1093/nar/gkac920

30. Liao H, Li H, Duan CS, Zhou XY, An XL, Zhu YG, Su JQ. 2022. Metagenomic and viromic analysis reveal the anthropogenic impacts on the plasmid and phage borne transferable resistome in soil. Environ Int 170:107595. https://doi.org/10.1016/j.envint.2022.107595

31. Ras V, Carvajal-López P, Gopalasingam P, Matimba A, Chauke PA, Mulder N, Guerfali F, Del Angel VD, Reyes A, Oliveira G, De Las Rivas J, Cristancho M. 2021. Challenges and considerations for delivering bioinformatics training in LMICs: perspectives from Pan-African and Latin American bioinformatics networks. Front Educ 6:710971. https://doi.org/10.3389/feduc.2021.710971

32. Tewhey R, Nakano M, Wang X, Pabón-Peña C, Novak B, Giuffre A, Lin E, Happe S, Roberts DN, LeProust EM, Topol EJ, Harismendy O, Frazer KA. 2009. Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol 10:1–13. https://doi.org/10.1186/gb-2009-10-10-r116

33. Gaudin M, Desnues C. 2018. Hybrid capture-based next generation sequencing and its application to human infectious diseases. Front Microbiol 9:2924. https://doi.org/10.3389/fmicb.2018.02924

34. Guitor AK, Raphenya AR, Klunk J, Kuch M, Alcock B, Surette MG, McArthur AG, Poinar HN, Wright GD. 2019. Capturing the resistome: a targeted capture method to reveal antibiotic resistance determinants in metagenomes. Antimicrob Agents Chemother 64. https://doi.org/10.1128/AAC.01324-19

35. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen A-LV, Cheng AA, Liu S, et al. 2020. CARD 2020: antibiotic resistome surveillance with the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res 48:D517–D525. https://doi.org/10.1093/nar/gkz935

36. Salverda MLM, De Visser JAGM, Barlow M. 2010. Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. FEMS Microbiol Rev 34:1015–1036. https://doi.org/10.1111/j.1574-6976.2010.00222.x

37. Pawlowski AC, Stogios PJ, Koteva K, Skarina T, Evdokimova E, Savchenko A, Wright GD. 2018. The evolution of substrate discrimination in macrolide antibiotic resistance enzymes. Nat Commun 9:112. https://doi.org/10.1038/s41467-017-02680-0

38. Daly M, Villa L, Pezzella C, Fanning S, Carattoli A. 2005. Comparison of multidrug resistance gene regions between two geographically unrelated *Salmonella* serotypes. J Antimicrob Chemother 55:558–561. https://doi.org/10.1093/jac/dki015

39. Scholz P, Haring V, Wittmann-Liebold B, Ashman K, Bagdasarian M, Scherzinger E. 1989. Complete nucleotide sequence and gene organization of the broad-host-range plasmid RSF1010. Gene 75:271–288. https://doi.org/10.1016/0378-1119(89)90273-4

40. De Oliveira DMP, Forde BM, Kidd TJ, Harris PNA, Schembri MA, Beatson SA, Paterson DL, Walker MJ. 2020. Antimicrobial resistance in ESKAPE pathogen. Clin Microbiol Rev 33:e00181-19. https://doi.org/10.1128/CMR.00181-19

41. Pendleton JN, Gorman SP, Gilmore BF. 2013. Clinical relevance of the ESKAPE pathogens. Expert Rev Anti Infect Ther 11:297–308. https://doi.org/10.1586/eri.13.12

42. Zhang Z, Zhang Q, Wang T, Xu N, Lu T, Hong W, Penuelas J, Gillings M, Wang M, Gao W, Qian H. 2022. Assessment of global health risk of antibiotic resistance genes. Nat Commun 13:1553. https://doi.org/10.1038/s41467-022-29283-8

43. Zhang A-N, Gaston JM, Dai CL, Zhao S, Poyet M, Groussin M, Yin X, Li L-G, van Loosdrecht MCM, Topp E, Gillings MR, Hanage WP, Tiedje JM, Moniz K, Alm EJ, Zhang T. 2021. An omics-based framework for assessing the health risk of antimicrobial resistance genes. Nat Commun 12:4765. https://doi.org/10.1038/s41467-021-25096-3

44. Campana MG. 2018. BaitsTools: software for hybridization capture bait design. Mol Ecol Resour 18:356–361. https://doi.org/10.1111/1755-0998.12721

45. Bar-On YM, Phillips R, Milo R. 2018. The biomass distribution on Earth. Proc Natl Acad Sci U S A 115:6506–6511. https://doi.org/10.1073/pnas.1711842115

46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

47. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. Nature New Biol 585:357–362. https://doi.org/10.1038/s41586-020-2649-2

48. Conrad T, Plumbom I, Alcobendas M, Vidal R, Sauer S. 2020. Maximizing transcription of nucleic acids with efficient T7 promoters. Commun Biol 3:439. https://doi.org/10.1038/s42003-020-01167-x

49. Wickham H. 2016. Ggplot2: elegant graphics for data analysis springer-verlag New York

50. Clausen PTLC, Aarestrup FM, Lund O. 2018. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics 19:307. https://doi.org/10.1186/s12859-018-2336-6

51. Tange O. 2011. GNU parallel: the command-line power tool. ;login

52. McKinney W. 2010. Data structures for statistical computing in python. Python in Science Conference; Austin, Texas, p 445. https://doi.org/10.25080/Majora-92bf1922-00a

53. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423. https://doi.org/10.1093/bioinformatics/btp163

54. PedregosaF, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V. 2011. Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830.

55. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN, Doshi S, Courtot M, Lo R, Williams LE, Frye JG, Elsayegh T, Sardar D, Westman EL, Pawlowski AC, Johnson TA, Brinkman FSL, Wright GD, McArthur AG. 2017. CARD 2017: expansion and model-centric curation of the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res 45:D566–D573. https://doi.org/10.1093/nar/gkw1004

56. Lanza VF, Baquero F, Martínez JL, Ramos-Ruíz R, González-Zorn B, Andremont A, Sánchez-Valenzuela A, Ehrlich SD, Kennedy S, Ruppé E, van Schaik W, Willems RJ, de la Cruz F, Coque TM. 2018. In-depth resistome analysis by targeted metagenomics. Microbiome 6:11. https://doi.org/10.1186/s40168-017-0387-y

57. Pehrsson EC, Tsukayama P, Patel S, Mejía-Bautista M, Sosa-Soto G, Navarrete KM, Calderon M, Cabrera L, Hoyos-Arango W, Bertoli MT, Berg DE, Gilman RH, Dantas G. 2016. Interconnected microbiomes and resistomes in low-income human habitats. Nature New Biol 533:212–216. https://doi.org/10.1038/nature17672

58. Lee K, Kim D-W, Lee D-H, Kim Y-S, Bu J-H, Cha J-H, Thawng CN, Hwang E-M, Seong HJ, Sul WJ, Wellington EMH, Quince C, Cha C-J. 2020. Mobile resistome of human gut and pathogen drives anthropogenic bloom of antibiotic resistance. Microbiome 8:2. https://doi.org/10.1186/s40168-019-0774-7

59. Wong KL, Liu J. 2021. Factors and methods to modulate DNA hybridization kinetics. Biotechnol J 16:2000338. https://doi.org/10.1002/biot.202000338

60. Ouldridge TE, Sulc P, Romano F, Doye JPK, Louis AA. 2013. DNA hybridization kinetics: zippering, internal displacement and sequence dependence. Nucleic Acids Res 41:8886–8895. https://doi.org/10.1093/nar/gkt687