Capturing the Resistome: A targeted capture method to reveal antibiotic resistance determinants in metagenomes

Allison K. Guitor, Amogelang R. Raphenya, Jennifer Klunk, Melanie Kuch, Brian Alcock, Michael G. Surette, Andrew G. McArthur, Hendrik N. Poinar, Gerard D. Wright

1 David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario, Canada
2 Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada
3 Department of Biochemistry and Biomedical Science, McMaster University, Hamilton, Ontario, Canada
4 McMaster Ancient DNA Centre, Department of Anthropology and Biochemistry, McMaster University, Hamilton, Ontario, Canada
Abstract

The identification and association of the nucleotide sequences encoding antibiotic resistance elements is critical to improve surveillance and monitor trends in antibiotic resistance. Current methods to study antibiotic resistance in various environments rely on extensive deep sequencing or laborious culturing of fastidious organisms, which are both heavily time-consuming operations. An accurate and sensitive method to identify both rare and common resistance elements in complex metagenomic samples is needed. Referencing the Comprehensive Antibiotic Resistance Database, we designed a set of 37,826 probes to specifically target over 2000 nucleotide sequences associated with antibiotic resistance in clinically relevant bacteria. Testing of this probeset on DNA libraries generated from multi-drug resistant bacteria to selectively capture resistance genes reproducibly produced higher reads on-target at greater length of coverage when compared to shotgun sequencing. We also identified additional resistance gene sequences from human gut microbiome samples that sequencing alone was not able to detect. Our method to capture the resistome enables sensitive gene detection in diverse environments where antibiotic resistance represents less than 0.1% of the metagenome.

Introduction

Antibiotic resistance (AMR) is one of the most pressing challenges of the 21st Century that poses a threat to modern medicine and food security (1). The challenge of AMR is amplified by the movement of genes between bacteria coupled with the movement of people and goods across the planet (2–4). One of the gaps to address the antibiotic crisis is a lack of suitable tools to catalogue the complete resistome (the entire AMR gene contingent) in various environments and associated microbiomes. Detecting the resistome of an individual bacterium, a microbiome, and other environmental settings (sediment, hospitals, etc.) will aid in tracking the spread of resistance and monitoring the emergence of new resistance alleles associated with the use of antibiotics or other bioactive compounds (5–10). This information can guide antibiotic use in addition to informing stewardship programs and public health decisions.
Profiling the resistomes of bacteria that are culturable is reasonably straightforward using whole genome sequencing followed by analysis using algorithms such as the Resistance Gene Identifier in the Comprehensive Antibiotic Resistance Database (CARD) (11). In metagenomes, where resistance determinants are relatively rare, deep sequencing requiring millions of sequencing reads followed by careful filtering is needed. This resource intensive strategy can be alleviated by targeted detection of selected genes, e.g. via polymerase chain reaction (PCR), microarrays, or CRISPR/Cas9-based methods (12–16). However, such highly targeted approaches suffer from the fact that they are rarely comprehensive and generally cannot account for the continual emergence of gene variants and/or completely novel mechanisms (17–19).

A more appropriate approach for the identification of resistomes in metagenomes is the use of a probe and capture strategy (20). Using this strategy we and others have captured, sequenced, and reconstructed human mitochondrial sequences as well as the genomes of infectious agents and extinct species from various environments including highly degraded archeological and historical samples (21–26). In a probe and capture experiment, target RNA ‘baits’ are designed to be complementary to target DNA sequences of interest. Synthesized probes are biotin-labelled and are incubated with the DNA from metagenomic or genomic libraries, where they hybridize to related sequences (Figure 1 a, b). Targets are captured using streptavidin-coated magnetic bead separation, then reactions are pooled and sequenced on a next-generation sequencing (NGS) platform (Figure 1 c, d, e). This strategy offers significant advantages for the sampling of resistomes in a variety of environments where resistance genes are generally rare and genetically diverse. Indeed, recently this approach has been explored for resistance gene capture by other groups (27–29). However, these accounts target many other genes that are not rigorously associated with resistance, increasing the sequencing cost and the opportunity for false positive gene identification.

Here we chronicle our targeted method for the analysis of antibiotic resistomes. We based our probeset design on stringently curated AMR gene (ARG) sequences from the CARD (v1.0.1 - 2015), tiled across ARG sequences, combined with rigorous analysis to suppress off-target hybridization. This design...
enables a more cost-effective and sensitive method to sample the known resistance gene landscape.

We test the efficacy of this probeset and our strategy using both a panel of pathogenic bacteria with known resistance genotypes, as well as uncharacterized human metagenomic stool samples. Our method demonstrates the superior design and methodology of the approach readily applicable to both clinical and non-clinical settings.

**Results**

**Design and characterization of resistance gene probes**

A set of 80-mer nucleotide probes were custom designed and synthesized through the myBaits® platform (Arbor Biosciences, Ann Arbor, Michigan). The probes (37,826) span the protein homolog model of curated ARGs from CARD and represent nucleotide sequences (2021) that are well-characterized in the literature. Resistance conferred through single point mutations (SNPs) (e.g. sequences contained in the protein variant model in CARD) in chromosomal metabolic genes (including DNA gyrase (gyrA) mutations associates with fluoroquinolone resistance and RNA polymerase subunit (rpoB) mutations associated with rifampin resistance) was purposefully not included in our design. 78.03% of genes targeted by our probes mirror the breakdown in CARD, dominated by antibiotic inactivation mechanisms and by the beta-lactamases (Figure 2). The majority of the probes (24,767) target a single gene and the remainder range to a maximum of 211 genes (average 5.96 genes) due to sequence conservation within gene families (Supplementary Figure 1A). For example, a single probe initially designed to target 80 nucleotides of the beta-lactamase gene bla\text{SHV-52} also targets an additional 208 genes including other members of the SHV, LEN, and OKP-A/B beta-lactamases due to homology between these nucleotide sequences within AMR gene families. The combination of overlap in the utility of some 80-mer probes and partial hybridization can allow for probes to target sequences that are divergent from their reference sequences and thus identify new alleles at the SNPS level up to 15% divergent.

At the individual determinant level, the number of probes per gene (average 105 probes per gene, range: 1 – 309) and length coverage of a gene (average 96.20%, range: 3.17% - 100%) varies.
The majority of targeted genes (2004/2021; 99.16%) are covered by at least 10 or more probes (Supplementary Figure 1B). Members of the beta-lactamase families (bla\text{CTX-M}, bla\text{TEM}, bla\text{OXA}, bla\text{GES}, bla\text{SHV}) are among the genes with the highest probe coverage. The majority of genes (1970/2021) have greater than 80% length coverage by probes, 26 genes have less than 50%, and only 1 (\textit{mexW}, ARO:3003031) has less than 5% length coverage by probes (Supplementary Figure 1C). Only 28 sequences from CARD have no probe coverage due to filtering of candidate probes during the design. Overall this probeset targets ~1.77 megabases of antibiotic resistance nucleotide sequence and greater than 83% of the nucleotide sequences curated in CARD. Additional metrics of the probeset are given in Supplementary Figures 1D-H.

**ARG enrichment from bacterial genomes with a range of antibiotic resistance determinants**

To characterize the sensitivity and selectivity of this probeset, we conducted a series of control experiments using a panel of sequenced multi-drug resistant Gram-positive and Gram-negative bacteria. The proportion of the genomes targeted by our probeset ranged from 0.21 – 0.97%, consisting of 13 to 65 ARGs, representing 102 unique genes among the isolates tested (Supplementary Table 1). Genomic DNA from four different species was tested individually via enrichment on two different library preparations (NEBNext Ultra II versus modified Meyer and Kircher) of varying insert sizes (average library fragment sizes range from 396 to 1257); referred to as Trial 1 and Trial 2 hereafter (Supplementary Table 2). Our enrichment approach is insensitive and tractable to different insert sizes as there is a strong correlation between read count on targeted regions for bacterial genomes enriched individually among the two trials (Pearson correlation 0.811 – 0.975) (Supplementary Table 3, Supplementary Figure 2).

This probeset is selective for regions associated with antibiotic resistance in these isolates given that over 90% of reads mapped to the respective draft bacterial genomes and the majority (higher than 85% in all cases) of reads mapped to the small proportion (<1%) of the genome associated with resistance (Supplementary Table 3; Figure 3A). We successfully captured 100% of the targeted genes in both library preparation methods with at least 10 reads and with 100% length coverage for the four species of bacteria.
tested (Supplementary Table 3). This represents a sensitivity ranging from 0.21% to 0.97% of the total DNA in these samples with successful enrichment of regions as small as 97 bp (mexW in Pseudomonas aeruginosa C0060 with a probe coverage of 2 had greater than 10 reads in both trials) and 80 bp (crp in Klebsiella pneumoniae C0050 had greater than 100 reads in both trials). Other genes that have low probe length coverage include mdtA (22.4% coverage by 11 probes) in Escherichia coli C0002 which still retained over 100 reads in both trials and a 140 bp region of aad(6) (16.8% coverage by 4 probes) in Staphylococcus aureus C0018 that was recovered with over 1000 reads in both trials.

Successful enrichment of ARGs in mock metagenomes

Genomic DNA from multiple bacteria was pooled at varying ratios of 4 or 8 isolates, with some bacteria representing less than 10% of the total mock metagenome (Supplementary Table 4). In 28/32 enrichments, 80% or more of the sequencing reads mapped to probe-targeted regions within the individual bacterial genome regardless of pooling ratios (Supplementary Table 5; Figure 3A). The one exception is Trial 1 Pool 2 (enriched), where on-target mapping was not as effective (~70%); nevertheless, even this trial remained over 50-fold better than the shotgun samples (Supplementary Table 5). In all shotgun samples, the percentage of reads on-target never exceeded 5% and in 31/32 cases was less than 2% of the total sequencing data (Supplementary Table 5, Figure 3A).

At the isolate level, the percentage of the mock metagenome that is represented by probe-targeted regions in an individual isolate ranged from 0.0015 to 0.63% of the total DNA (Supplementary Table 4, Supplementary Figure 3). In 21/32 enriched cases, over 90% of probe-targeted regions were captured by 10 reads or more (Supplementary Table 5; Figure 3B). In contrast, none of the shotgun sequenced samples recovered more than 80% of probe-targeted regions with at least 10 reads. The cases in which enrichment underperformed are associated with two species in particular: K. pneumoniae and P. aeruginosa (Supplementary Table 5; Figure 3B). We define the sensitivity of detection of AMR for a given isolate as the percentage of total DNA represented by probe-targeted regions of a given genome at which greater than 90% of probe-targeted regions were recovered with at least 10 reads. These values...
range from 0.033% for *S. aureus* C0018 to 0.11% for *P. aeruginosa* C0060 (Supplementary Figure 3).

With these bacterial species tested, our probeset can successfully capture the resistome of these isolates that represents less than 0.1% of the total DNA and even less at the individual gene level.

**Enrichment from mock metagenomes exceeds shotgun sequencing**

We recovered significantly more targeted genes with at least 1, 10, or 100 reads mapping (mapping quality \(\geq 41\) and length \(\geq 40\)) compared to shotgun sequencing (Figure 3B; Supplementary Table 5). Furthermore, the average percent coverage of probe-targeted regions with at least 1, 10, or 100 reads in all isolates enriched individually or in pools is always higher than in the shotgun samples and ranges from 1.05- to 18.3-fold greater (Figure 3C, Supplementary Table 5). For all genomes in all pooled libraries across both trials, the average normalized read count and depth of reads on probe-targeted ARGs from enriched libraries is over 50 times (57.09 – 25683.42) higher than from its shotgun control (Supplementary Table 5). In 31/32 cases, the fold-increase in read counts exceeded two orders of magnitude and was over four for some probe-targeted regions (Supplementary Table 5). The one case that did not conform (from Trial 1 Pool 2, see above) reflects a minor and non-reproducible variability in the quality of the capture for unknown reasons. Nonetheless, there is a clear distinction between the shotgun and enriched samples with the enriched data showing a more consistent agreement between normalized read counts per probe-targeted region (Figure 4). A similar trend is observed when the raw read counts for each sample are used (Supplementary Figure 4).

**ARG analysis of a human gut metagenome**

In order to test the efficacy and reproducibility of our enrichment in more complex samples, we performed enrichments on replicates from metagenomic libraries with DNA isolated from a healthy individual’s stool sample. We compared experiments with traditional shotgun sequencing whereby selected libraries were sequenced to a depth of over 3.5 million paired reads (Supplementary Table 2). We included a series of positive control enrichments with genomic DNA from *E. coli* C0002 that was used previously for enrichment with the mock metagenomes. In all cases, we identified the same genes with a
consistent number of reads mapping among these replicate enrichments (when subsampled to equal
depths among sets) proving reproducibility regardless of probe and library ratio (Supplementary Table 6;
Supplementary Figure 5). Within each set, we found an excellent correlation with previous results seen
with E. coli C0002 (Pearson correlations: >0.923 for all pairs in Set 1, >0.924 for Set 2, >0.901 for Set 3)
(Supplementary Figure 5).

Across the enriched gut microbiome samples, with the full number of reads and no filters, on-
average 50.69% of reads map to CARD and 68 genes were identified with at least 10 reads, compared to
0.03% mapping and 32 genes identified in the shotgun libraries (Supplementary Figure 6A, B;
Supplementary Table 7). We found significantly more genes with at least 1, 10, and 100 reads from each
enriched sample as compared to the shotgun samples and that the average percent coverage of a gene by
reads in the enriched samples is 1.5-fold higher (Supplementary Figure 6 B, C). When subsampled to the
same depth as their enriched pairs (between 22,324 and 149,320 reads), we identified on average 1
(range: 0 – 2) antibiotic resistance determinant with at least 10 reads after filtering in the shotgun samples
making comparisons at this level unrealistic (Supplementary Table 8). Conversely, when subsampled to
the depth of the lowest enriched sample (22,324 reads), we identified on average 28 ARGs in the enriched
libraries post-filtering with at least 10 reads (Supplementary Table 8).

High fold-enrichment of ARGs from human stool

We combined the read counts for genes with at least 10 reads that passed the chosen filters within
each set to compare between probe and library ratios in subsampled and full read samples through both
enrichment and shotgun sequencing. With the full number of reads, 24/70 (34.28%) of genes detected
overlap among all enriched libraries (n = 27), while we identified 16 genes of a total 32 (50.00%) in all
the shotgun libraries (n = 6, Supplementary Table 7, 9). When subsampled to the lowest enriched read
coverage (22,324 reads), there are no genes that overlap between all six shotgun libraries, while 13/47
(27.66%) of genes overlap across all 27 enriched libraries (Supplementary Table 10). Comparing among
subsampled enriched libraries (22,324 reads), the majority (31/34) of genes missing in at least one sample
are those with on average less than twenty reads across the 27 libraries (Supplementary Tables 10;
Supplementary Figure 7). The order of genes with higher read counts (i.e. higher abundance and gene copy number) is consistent among enriched and shotgun samples and there is a more significant discrepancy between reads associated with lower abundance genes (Supplementary Figure 7, 8). Thus, enrichment, in the same way as shotgun sequencing, does not in some way bias the prevalence of rank order of AMR in these samples. Finally, both methods resulted in excellent correlation among technical replicates individually (Pearson correlation 0.871 for shotgun and 0.972 for enriched; Supplementary Figure 7, 8).

We found that enrichment exceeded shotgun sequencing by identifying more unique antibiotic resistance genes at much lower sequencing depths. The enriched samples provided a more diverse representation of ARGs at less than 100,000 paired reads compared to over 5 million reads in the shotgun samples (Supplementary Figure 8, 9). With the full number of reads in both methods (between 66- and 389-fold more in the shotgun samples than the enriched samples), the average fold-enrichment is >600-fold and there are still 18 to 50 fewer genes in the shotgun samples (Figure 5A; Table 1). In most cases, there are only a few genes found via shotgun that are missing in the enriched paired sample (between 9 and 15; 22 unique genes). Only between 1 to 5 of these genes (7 total unique) in each sample is predicted to be targeted by probes (Table 1). Of these, only one, novA, is missing from all enriched samples but is present in all shotgun samples with >10 reads, mapping quality ≥11 and percent length coverage by reads ≥10%. The other 6 genes (macB, vanRG, vanSG, smeE, cfxA6, cepA) are found in only a few shotgun samples with less than 30 reads and less than 20% read length coverage on average (Table 1; Supplementary Table 13). When combined, the enriched libraries cluster separately from the shotgun libraries with a stronger correlation (0.9957 compared to 0.8712 for the shotgun libraries; Supplementary Figure 8).

We then compared the overlap between all 27 enriched samples and the six shotgun-sequenced libraries and included genes found through shotgun without any probes mapping. We found a total of 89 genes with at least 10 reads between all libraries of which, 13 are overlapping between methods, 57 are unique to the enriched libraries, and 19 are unique to the shotgun libraries (Figure 5B; Supplementary Figure 7).
Table 13). Of the 19 genes not found in any enriched library, only 4 are predicted to be targeted by probes, while the remaining were not in CARD when the probes were initially designed (8) or had probes that were removed during design and filtering (7). Of the four genes with predicted probes, cfxA6 is present in all enriched samples but was filtered out by mapping quality; vanSG is only present in 2/6 shotgun samples at less than 20% gene length coverage by reads; cepA is found in enriched samples but at less than 10 reads; finally, we identified novA in all shotgun samples but in only a few enriched samples at less than 10 reads and less than 10% read length coverage. Despite the few (4) genes that are missing from the enriched samples, even a 200-fold greater sequencing depth of our shotgun libraries could not match our enrichment data (Supplementary Figure 9).

Negative control results

To track and measure the contamination in our lab, commercial kits, environment and reagents, we included negative controls of a blank DNA extraction and negative reagent controls in enrichment that we processed identically to our samples used in Phase 1 and 2. For Phase 1 in both Trial 1 and Trial 2, we found a negligible amount of library DNA in the blank after enrichment and very few of the sequenced reads were associated with the indexes used for the blank library (between 2.46% and 8.96% of sequenced reads; Supplementary Table 2, 11). Only the blank samples from Phase 1 Trial 1 and Phase 2 Set 2 resulted in genes with at least 10 reads mapping (10 and 19 respectively; Supplementary Table 12).

Discussion

Increased interest in targeted capture approaches has resulted in the design of probesets for the detection of viruses, bacteria, and more recently, antibiotic resistance elements (26–29). Although our study is not the first to employ targeted capture for antibiotic resistance genes, we have focused on a rigorous probe design that includes choosing an appropriate reference database, robust probeset validation, and experimental considerations for enrichment including reduced input library and probe concentrations (25, 30–33). Our probe design and application of in-solution targeted capture ultimately
results in a cost-effective alternative to shotgun sequencing for identifying antibiotic resistance genes in complex environmental and clinical metagenomes.

**Reference database for probe design and analysis**

CARD was chosen as the reference database for our probe design (v.1.0.1) and analysis (v.3.0.0) due to its rigorous curation of antibiotic resistance determinants. We excluded some genes (e.g., *gyrA*, EF-Tu genes, etc.) that are likely to be found as homologs across many families of bacteria and would likely have overwhelmed the probeset and sequencing effort with abundant, non-mutant antibiotic susceptible alleles. Instead, we chose CARD’s protein homolog model (v.1.0.1, n = 2010) to focus our approach on genes that are likely to be acquired (i.e. associated with mobile genetic elements) and those that are unique to individual families of bacteria. Therefore, although we are unable to detect resistance conferred by SNPs in chromosomal metabolic genes, our probeset is capable of capturing the vast majority of resistance elements and those that are at a higher risk of being mobile. In future probe designs, the protein variant model of CARD (v.1.0.1, n = 77; v.3.0.0, n = 141) can be targeted using probes specific to the regions of a gene associated with a given set of SNPs, but they will need to be carefully tested in-silico to ensure that they do not enrich unintended targets. Given that in certain populations (e.g., metagenomes) these variant sequences may be less abundant than their susceptible counterparts, careful and rigorous analysis will need to be implemented to identify the relevant variants (i.e., RGI developments).

To address our probeset’s compatibility with a frequently updated database, we chose a more recent version of CARD (v.3.0.0 n = 2238) for comparative analysis to our bait set designed in 2015. Since the design of our probes against v1.0.1 of CARD, the database has been updated and includes 264 additional genes. Despite these changes, our probeset targets the majority (2021/2238) of known antibiotic resistance sequences from CARD v3.0.0. In reality, the probes should target sequences with up to 15% nucleotide divergence from a reference, suggesting wider applicability and target capacity towards newly characterized members of AMR gene families, which often only differ from other members by a few nucleotides. Of the 264 genes added to CARD v3.0.0, our existing probes capture 75 of these genes that are sufficiently similar to other targeted members of the same AMR gene family (e.g., aminoglycoside...
acetyltransferases, chloramphenicol acetyltransferases, beta-lactamases (bla\textsubscript{ACT}, bla\textsubscript{CARB}, bla\textsubscript{CMY}, bla\textsubscript{LEN}, bla\textsubscript{NDM}, bla\textsubscript{OXA}, bla\textsubscript{PDC}, bla\textsubscript{SHV}, bla\textsubscript{TEM}, bla\textsubscript{VEB}). Of the remaining genes, 60 were newly identified sequences since 2015, and the others, although mentioned in the literature prior to 2015, were added due to increased efforts of curation of CARD.

Other approaches targeting ARGs have included species identifiers, plasmid markers, and biocide or metal resistance (27–29). These probesets range in target capacity from 5557 genes (3.34 Mb) (28) to over 78,600 genes (88.13 Mb) (27) and comprise up to 4 million probes (29). Other strategies involve designing one probe per gene, tiling probes across a gene without overlap (1X coverage), or inter-probe distances of up to 121 nucleotides (28, 29). Our approach is more conservative in probe design (1.77 Mb for 2021 genes), but the dense tiling allows for more probes per gene (99.16% of genes with greater than 10 probes) and increased depth of probe coverage (9.47X average). We believe the design approach increases specificity, sensitivity, and the likelihood of capturing rare DNA molecules common in complex metagenomes (34). We also performed extensive filtering of candidate probes against the human genome, other eukaryote, archaeal, and weakly matching bacterial sequences to provide a probeset that is bacterial ARG specific and avoids off-target hybridization. Focusing on one highly curated database of antibiotic resistance determinants (CARD) increases the likelihood of capturing bona fide sequences that are associated with known resistance and reduces the overall cost of the probeset and sequencing effort. When updates to CARD are released, or if additional markers are of interest, probes can easily be designed and added to the existing probeset.

**Experimental considerations in targeted capture methods**

For our trials we tested significantly lower inputs (25 ng to 400 ng) than recommended by the manufacturer (up to 2 µg of DNA for metagenomic samples) setting us apart from other targeted capture methods of AMR genes (27, 28). Others have looked at reducing the amount of input DNA from the manufacturer’s recommended amount of 3000 ng to 500 ng and saw no significant differences in results (35). Despite a 16-fold reduction in DNA input (25 ng vs the recommended 2000 ng), we saw no visible differences in the order of genes captured in the stool sample and normalized read counts were
comparable among different library and probe amounts, suggesting that our approach is robust to
tremendous fluctuations yet still identifies all antibiotic resistance genes in samples with low DNA yield
(e.g. clinical and environmental samples). Furthermore, a lower input concentration of probes also
reduces the cost per reaction.

**Reproducibility, sensitivity, and performance with clinical isolates**

The sensitivity of our probeset was tested using individual bacterial genomes and mock
metagenomes wherein the percentage of total DNA represented by probe-targeted antibiotic resistance
genes ranged from 0.0015% to 0.97%. A successful enrichment in our trials is considered when greater
than or equal to 90% of probe-targeted regions are captured with 10 or more reads. When tested
individually, enrichment was able to successfully capture all probe-targeted ARGs (100% with more than
10 reads) in the four bacterial species tested with >85% of sequenced reads mapping to the targeted
regions (<1 %) of the genome. With the mock metagenomes, the probe-targeted regions of each isolate
represented a smaller proportion and there were 11 cases in which enrichment was not successful under
the above criterion. In 7 instances, the given isolate represented less than 10% of the total pool and many
of the probe-targeted regions that are missing are short (<200 bp) with less than 5 probes designed.
(Supplementary Table 4). One particular predicted resistance gene that was not captured in 2 cases, ARO:
3002804 $\text{fosA2}$ in *K. pneumoniae* C0050, retains good probe coverage despite a low percent identity
(71.32%) to the CARD reference. The poor performance in enrichment may suggest the limit of sequence
similarity (>30%) that can be captured by probes designed against a single reference sequence. In
addition, the high GC content of certain genes in the *K. pneumoniae* isolates and of many regions of the
*P. aeruginosa* isolates (average 67% GC), likely reduced the capture efficiency in the more complex
pooled samples resulting in less than 10 reads on-targeted genes. The conditions of hybridization may
need to be further optimized for higher GC content targets. Regardless of this limitation, the enriched data
provides significantly more read coverage for antibiotic resistance genes at a lower depth of sequencing
as compared to the shotgun sequencing of these mock metagenomes.
Standardization and controls in metagenomics

Standardization (including reproducibility) in enrichment studies remains sorely lacking. In this study, we attempted to reduce bias and assess enrichment by using the same DNA extract, library preparations, and enrichment in triplicate. Even among replicate libraries and shotgun sequencing runs, the differences in the number of genes identified at various sequence depths highlights the inherent variability in metagenomics (Supplementary Figure 8). The positive control (E. coli C0002), processed alongside other samples, ensured our methodology and probes were performing optimally at the time of hybridization. We also introduced negative controls to measure the extent of exogenous DNA contamination that is ubiquitous in all laboratory settings and reagents (36, 37). Between 86.07 and 100% of the sequenced reads from our negative controls had corresponding index sequences from experimental samples, suggesting that DNA exchange among samples during enrichment or cross-contamination is the primary concern in our method (Supplementary Table 2, 11). Notably, the genes identified in the blank results not arising from cross-contamination and also found in the enriched and shotgun results are commonly associated with bacteria identified in negative controls in microbiome studies (mainly Escherichia coli) and encode efflux systems or other intrinsic resistance determinants (mdtEFHOP, emrKY, cpxA, acrDEFS, pmrF, eptA, tolC). The two genes that were unique to the blank results (drfA17 had 11 reads covering 85.86%; aph(3'')-Ib 16 reads with 57.46% coverage) are associated with mobile genetic elements in Enterobacteriaceae and the latter has been previously associated with laboratory reagent contamination (38, 39). Despite standard methods to control for contamination (i.e. filter pipette tips, PCR cabinets, and sterile DNA/RNA-free consumables), we still find limited contamination likely stemming from reagents and/or the surrounding laboratory environment, further highlighting the importance of negative controls in all targeted capture experiments and meticulous reporting and publishing of a laboratory based ‘resistome’ (Supplementary Table 6) (36, 37, 40).

Enrichment in the gut microbiome

Our enrichment of resistance genes in the human gut microbiome samples resulted in a higher average percentage on-target (50.69%) when compared to other published capture-based methods.
348 30.26% (20.27 – 41.83%) (27), and a median of 15.8% (0.28% - 68.2%) (28) highlighting the increased
349 specificity of our probe design. Overall, our probeset and method identified a greater diversity of
350 antibiotic resistance genes in the human gut microbiome despite having been sequenced at 66 – 389-fold
351 lower depth when compared to their shotgun sequenced correlate. With a reduced depth of sequencing, it
352 is evident that enrichment offers more valuable information in both the number of genes with reads as
353 well as the depth and breadth of coverage of those genes (Figure 5).

Although shotgun sequencing can provide additional information on other functions and genes of
355 interest, our targeted capture provides a more robust, reproducible profile of antibiotic resistance genes
356 from metagenomes at a fraction of the sequencing cost. Only a few genes were absent in the enriched
357 libraries when compared to the shotgun libraries. In the case of novA, which is 70.51% GC, there is a gap
358 in tiling of probes across the gene and perhaps hybridization conditions were not sufficient to capture this
359 gene by our method. Additional probes or more dense tiling along high GC content (>65%) sequences
360 GC) sensor from vancomycin resistance gene clusters, was covered by less than 20 reads in the shotgun
361 samples, suggesting a very low abundance in the metagenome. Finally, the beta-lactamase genes cepA
362 and cfxA6 had been excluded from the enriched results after filtering due to low mapping quality or less
363 than 10 reads. The low mapping quality suggests that reads are mapping to other beta-lactamase genes in
364 the reference database.

All current methods to detect antibiotic resistance genes have limitations. Culturing, although
366 time-consuming, remains the standard for diagnosing infections through the identification of both the
367 pathogen and its susceptibility to a panel of antibiotics. Other biochemical techniques have been
368 developed but are often organism-specific and require additional assays for confirming ARGs (41). When
369 studying the microbiome and the resistome of various environments, a culture-based approach is not
370 feasible and thus high-throughput methods are needed (19). Sequencing-based approaches (i.e. PCR-
371 based, microarray-based, and in-solution targeted capture) and quantitative PCR methods offer selective
372 and sensitive means to identify a larger contingent of antibiotic resistance genes but can be (or are

15
designed to be) heavily biased or selective. While PCR is highly sensitive, many panels for AMR target
only a range of between 200 to 400 genes (42). As we have shown here, probe-based hybridization
methods enable the detection of over 2000 ARGs in a single assay.

Compared to the other probesets designed for AMR (27–29), ours offers a highly curated specific
set of probes, with high coverage of ARGs, and works exceptionally well on low input samples. We have
also included crucial controls to validate our findings. Where shotgun sequencing requires millions of
reads to detect a few antibiotic resistance genes, we have shown that targeted capture can detect the same
genes and more with ~50-fold less sequencing effort. A reduced amount of sequencing allows more
samples to be processed per individual sequencing run, reducing sequencing costs overall and increasing
throughput. One limitation to targeted approaches is that the probe design relies on known reference
sequences, while shotgun sequencing can reveal additional information not captured by the probes but at
an added cost (depth). All sequencing-based methods are limited in the inability to characterize
completely novel antibiotic resistance determinants, whereas a functional metagenomics approach is ideal
in this regard (19).

In conclusion, we have rigorously measured the performance of our probe design and methods to
satisfy many of the parameters routinely discussed in targeted capture (43). The sensitivity and specificity
of our probeset is evident by the consistently high percentage of reads on-target and high recovery of
probe-targeted sequences representing <0.1% of the total DNA. Our approach results in the uniform
recovery of ARGs across bacterial genomes and is reproducible between library preparations. We believe
our targeted capture serves a critical role in the surveillance and detection of ARGs across complex
environmental settings, hospital and clinics. Profiling these resistomes will provide invaluable
information that can be used to target antibiotic and resistance inhibitor discovery while at the same time
keeping abreast of the rapidly shifting rise of local and global antibiotic resistance.
Methods

Nucleotide probe design and filtering to prevent off-target hybridization

Our reference for probe design was the protein homolog model of antibiotic resistance determinants (n = 2,129) from the CARD (version 1.0.1 released December 14, 2015) (11). Using PanArray (v1.0), we designed probes of 80 nucleotide length across all genes with a sliding window of 20 nucleotides and acceptance of 1 mismatch across probes (32). To prevent off-target hybridization between the probes and non-bacterial sequences, the candidate set of probe sequences (n = 38,980) was compared against the human reference genome and GenBank’s non-redundant nucleotide database through BLAST (blastn) (44, 45). Probes with high sequence similarity (>80%) and probes with high-scoring segment pairs (HSPs) greater than 50 nucleotides of a possible 80 were discarded (human n=158, eukaryotic n = 1617, viral n = 774 and archaeal n = 30). Probes with HSPs less than 50/80 nucleotides to bacterial sequences were additionally discarded, resulting in a set of 32,066 probes. The candidate list was further filtered to omit probes that had bacterial HSPs that were <95% identity, resulting in a candidate list of 21,911 probes.

Optimizing probe density and redundancy

Probe sequences, along with 1-100 nucleotide(s) upstream and downstream of the probe location on the target gene, were sent to Arbor Biosciences (Ann Arbor, MI) for probe design. These sequences are contained within the open reading frame of the target gene and allow probe sequences to be modified if needed (i.e. polynucleotides at the termini) ensuring desired probe coverage of target genes is attained. Additional 80 nucleotide probes were created across the candidate probe and flanking sequences at four times tiling density, resulting in 226,440 probes. Sequences with 99% identity over 87.5% length were collapsed using USEARCH (usearch -cluster_fast -query_cov 0.875 -target_cov 0.875 -id 0.99 -centroids) resulting in a set of 37,826 final probes (46). Filtering similar to as described above was performed against the human genome; no probes were found to be similar. Arbor Biosciences (Ann Arbor, MI) synthesized this final set of 37,826 80-nt biotinylated ssRNA probes through the custom myBaits® kit.
Probe assessment and predicted target genes

To predict the genes that can be targeted by the probes, a Bowtie2 (settings used: bowtie2 --end-to-end -N 1 '-L 32' -a) (47) alignment was performed to compare the set of 37,826 probe sequences to the 2,238 nucleotide reference sequences of the protein homolog models in CARD (version 3.0.0 released 2018-10-11). The alignment file was manipulated through samtools and bedtools to determine the number of instances that a probe mapped to a nucleotide sequence in CARD, the fraction of each gene sequence covered by probes (length coverage by probes), and the depth of coverage by probes of each gene (bedtools genomcov, bedtools coverage -mean) (48, 49). The GC content of probe sequences and nucleotide sequences in CARD was calculated using a Python3 script from https://gist.github.com/wdecoster/8204dba7e504725e5bb249ca77bb2788. Melting temperature (Tm) was determined using OligoArray function melt.pl (-n RNA, -t 65 -C 1.89e^-9) (50). We used Prism 8 for macOS (https://www.graphpad.com) to generate plots in Supplementary Figure 1.

Bacterial isolates, samples, and DNA extraction

Clinical bacterial isolates were obtained from the IIDR Clinical Isolate Collection which consists of isolates from the core clinical laboratory at Hamilton Health Sciences Centre (Supplementary Table 1). Genomic DNA was isolated from a cell pellet using the Invitrogen Purelink Genomic DNA kit (Carlsbad, CA). If DNA was not isolated the same day, we stored cell pellets at -80°C. While genomic DNA from all other isolates was only extracted once, DNA from a cell pellet of Pseudomonas aeruginosa C0060 was extracted additionally using a varied genomic lysis/binding buffer (30 mM EDTA, 30 mM Tris-HCl, 800 mM GuSCN, 5% Triton-X-100, 5% Tween-20, pH 8.0). We obtained a human stool sample from a healthy volunteer for the purpose of culturing the microbiome with consent (HiREB#5513-T). DNA was extracted the same day following a modified protocol as described in (51). Briefly, samples were bead beat, centrifuged, and the supernatant further processed using the MagMax Express 96-Deep Well Magnetic Particle Processor from Applied Biosystems (Foster City, CA) with the multi-sample kit (Life Technologies #4413022). DNA was stored at -20°C until used for library preparation.
Isolate genome sequencing

Library preparation for genome sequencing of the clinical bacterial genomes was completed by the McMaster Genomics Facility in the Farncombe Institute at McMaster University (Hamilton, ON) using the New England Biolabs (Ipswich, MA) Nextera DNA library preparation kit. Libraries were sequenced using an Illumina HiSeq 1500 or Illumina MiSeq v3 platform using V2 (2 x 250 bp) chemistry. Paired sequencing reads were processed through Trimmomatic v0.39 to remove adaptors, checked for quality using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and de novo assembled using SPAdes v 3.9.0 (52, 53). The Livermore Metagenomics Analysis Toolkit (LMAT) v 1.2.6 was used to identify the bacterial species and screen for contamination or mixed culture, while the Resistance Gene Identifier (RGI; version 4.2.2) from CARD was used on the SPAdes contigs to identify Perfect (100% match) and Strict (<100% match but within CARD similarity cut-offs) hits to CARD’s curated antibiotic resistance genes (54).

Trials for enrichment

We performed two phases of experiments, the first with genomic DNA from cultured multi-drug resistant bacteria (Phase 1) and the second with metagenomic DNA from a human stool sample (Phase 2). The two trials in Phase 1 differ in their library preparation methods as described below (the major difference being library fragment size by sonication). In both trials, we tested genomic DNA from isolates individually (Escherichia coli C0002, Pseudomonas aeruginosa C0060, Klebsiella pneumoniae C0050, and Staphylococcus aureus C0018) (Supplementary Table 1 and 3). In addition, varying nanogram amounts (based on absorbance; Thermo Fisher Nanodrop, Waltham, MA) of each genome were combined prior to library preparation to create “mock metagenomes” referred to as Pool 1 (C0002, C0018, C0050, C0060), Pool 2 (C0002, C0018, C0050, C0060), and Pool 3 (C0002, C0018, C0050, C0060, Klebsiella pneumoniae C0060, Staphylococcus aureus C0033, Escherichia coli C0094, Pseudomonas aeruginosa C0292). Amounts of each isolate in each Pool varied between trials (Supplementary Table 4). Phase 2 consists of 3 replicates referred to as Set 1, Set 2, and Set 3 wherein a DNA extract from one individual human stool sample was split evenly into each Set. From these aliquots, we generated 9 individually...
indexed sequencing libraries and performed capture with varying library and probe ratios (Supplementary Table 3). In all trials and sets, a blank DNA extract was carried throughout library preparation and enrichment, while an additional negative reagent control was introduced during enrichment.

**Library preparation for enrichment sequencing**

Library preparations were performed in a PCR clean hood, using bleached equipment, and UV-irradiated before use to prevent non-endogenous DNA contamination. Trial 1 used the NEBNext Ultra II DNA library preparation kit (New England Biolabs, Ipswich, MA) through the McMaster Genomics Facility. Based on absorbance and fluorometer values (QuantiFluor, Promega, Madison, WI), we sonicated approximately 1 microgram of individual bacterial genomic DNA or pools of genomic DNA to 600 base pairs (bp) and prepared dual-indexed libraries with a size selection for 500-600 bp inserts. Post-library quality and quantity verification was performed using a High Sensitivity DNA Kit for the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) and quantitative PCR using the KAPA SYBR Fast qPCR master mix for Bio-Rad machines (Roche Canada) using primers for the distal ends of Illumina adapters and the following cycling conditions: 1) 95°C for 3 min; 2) 95°C for 10 sec; 3) 60°C for 30 sec; 5) Repeat 2-3 for 30 cycles total; 6) 60°C for 5 min 7) 8 °C hold. We used Illumina’s PhiX control library (Illumina, San Diego, CA) as a standard for quantification.

In Trial 2, the same genomic DNA, except for *P. aeruginosa* C0060 which was re-isolated, was used for library construction through a modified protocol (Supplementary material) (55). Briefly, we performed blunt end repair, adapter ligation, a library size-selection, and indexing PCR on ~200 nanograms of sonicated DNA (250-300 bp). The McMaster Genomics Facility performed library quality control as described above.

**Library preparation from a human stool sample**

We divided one DNA extract from a donor stool sample into three 50 µL aliquots of approximately 3150 nanograms each (based on fluorometer QuantiFluor results). DNA was sonicated to 600 bp and split into 9 individual library reactions (350 ng in 5.55 µL). We prepared dual-indexes libraries (NEBNext Ultra II library kits, New England Biolabs, Ipswich, MA) with a size-selection for 20
700-800 bp library fragments and 6 (Set 1), 7 (Set 2), or 8 cycles (Set 3) of amplification. The McMaster Genomics Facility performed library quality control (Agilent Bioanalyzer 2100 and quantitative PCR as described above). We generated positive control libraries using *Escherichia coli* C0002 genomic DNA (40 ng of sonicated DNA) and a negative control with a blank DNA extract.

**Targeted capture of bacterial isolates**

We performed enrichments in a PCR clean hood, with a water bath, thermal cyclers and heat blocks located nearby. The probeset was provided by Arbor Biosciences (Ann Arbor, MI) and diluted with deionized water. For enrichment of bacterial genomes in Trial 1, we used 100 ng of probes and 100 ng of each library following the myBaits® Manual V3 (Arbor Biosciences, Ann Arbor, MI) at a hybridization temperature of 65°C for 16 hours (see supplementary methods for more details). After hybridization and capture with Dynabeads MyOne Streptavidin C1 beads (Thermo Fisher, Waltham, MA), the resulting enriched library was amplified through 30 cycles of PCR (cycling conditions in Supplementary materials) using the KAPA HiFi HotStart polymerase with library non-specific primers (Kapa Library Amplification Primer Mix (10X), Sigma-Aldrich, St. Louis, MO). A 2 µL aliquot of this library was amplified in an additional PCR reaction for 3 cycles (same conditions as above) and then purified. We performed the capture in Trial 2 the same as Trial 1 but applied 17 cycles of amplification post-capture (PCR conditions in Supplementary details). The McMaster Genomics Facility performed library quality control as described above. Libraries were pooled in equimolar amounts and sequenced to an average of 94,117 clusters by MiSeq V2 (2x250 bp reads). Pre-enrichment libraries for the “mock metagenomes” were sequenced on a separate MiSeq V2 (2x250 bp reads) run from the enriched libraries to an average of 93,195 clusters each. From both Trial 1 and Trial 2, negative controls of blank extractions carried through library preparation and enrichment were sequenced on separate individual MiSeq 2 x 250 bp runs. After de-multiplexing of the blank, all possible index combinations were retrieved to identify potential cross-contamination of libraries as well as exogenous bacterial contamination.
Targeted capture of the stool sample

Based on qPCR values and the average fragment sizes of each library generated from the human stool DNA extract, we combined varying nanogram amounts of library (50, 100, 200 ng) and probes (25, 50, 100, 200, 400 ng) for enrichment (Supplementary Table 3). Along with the Negative Control - blank library, we introduced additional negative controls during enrichment using dH₂O to replace the volume normally required for library input. We performed enrichment following the myBaits® Manual V4 (Arbor Biosciences, Ann Arbor, MI) at a hybridization temperature of 65°C for 24 hours. After hybridization and capture with Dynabeads (Thermo Fisher, Waltham, MA), the resulting enriched library was amplified through 14 cycles of PCR using the KAPA HiFi HotStart ReadyMix polymerase with library non-specific primers and the same conditions as above (Supplementary Enrichment methods). The resulting products were purified using KAPA Pure Beads at a 1X volume ratio and eluted in 10 mM Tris, pH 8.0. Purified libraries were quantified through qPCR using 10X SYBR Select Master Mix (Applied Biosystems, Foster City California) for BioRad Cfx machines, Illumina specific primers (10X primer mix from KAPA) and Illumina’s PhiX Control Library as a standard. Cycling conditions were as follows: 1) 50 °C for 2 min; 2) 95 °C for 2 min; 3) 95 °C for 15 sec; 4) 60 °C for 30 sec; Repeat 3 – 4 for 40 cycles total. We pooled enriched libraries in equimolar amounts based on qPCR values and the McMaster Metagenomic Sequencing facility performed library quality control as described above. Finally, we sequenced the enriched libraries (average of 97,286 clusters) and the pre-enrichment libraries (average of 5,325,185 clusters) by MiSeq V2 2x250 bp. The negative controls of blank extractions carried through library preparation and enrichment were sequenced on separate individual MiSeq 2 x 250 bp runs. After de-multiplexing, all possible index combinations were retrieved.

Analysis of the bacterial isolates sequencing data

In order to identify probe-targeted regions and coordinates that overlap with predicted resistance genes based on RGI results for the individual bacterial genomes, we aligned our probeset to the draft reference genome sequence using Bowtie2 version 2.3.4.1 (47). We used Skewer version 0.2.2 (skewer -m...
pe -q 25 -Q 25) to trim sequencing reads (enriched or shotgun), bbmap version 37.93 dedupe2.sh to remove duplicates, and mapped reads to the bacterial genomes using Bowtie2 version 2.3.4.1 (--very-sensitive-local, unique sites only) (https://sourceforge.net/projects/bbmap/; (47, 56)). Aligned reads were filtered based on mapping quality (≥ 41) and length (≥ 40 bp) using various tools: samtools version 1.4, bamtools version 2.4.1, and bedtools version 2.27.1 (48, 49, 57). We determined the number of reads mapping to the reference genome overall and the number of reads mapping within a predicted probe-targeted region using genomic coordinates and bedtools (intersectBed) (50). The percent length coverage and the average depth of coverage of each probe-targeted region with at least one read was determined using bedtools coverage (-counts, -mean and default function) (49). We normalized read counts by the number of reads mapping per kb of targeted region per total number of mapping reads to a particular genome. The number of genes with at least 1, 10, or 100 reads were counted and their percent length coverage by reads was determined.

**Analysis of stool sample sequencing data**

We processed the enriched and shotgun reads for the human stool sample in the same way as for the bacterial isolates. Subsampling of reads was performed using seqtk version 1.2-r94 (seqtk sample -s100; https://github.com/lh3/seqtk). We used the bwt feature in RGI (beta of version 5.0.0; http://github.com/arpcard/rgi) to map trimmed reads using Bowtie2 version 2.3.4.1 to the CARD (version 3.0.0) generating alignments and results without any filters (47). We parsed the gene mapping and allele mapping files to determine the number of genes in CARD with reads mapping (at least 1, 10, and 100 reads) under various filters. After plotting mapping quality for each read in every sample across the 3 sets, we chose an average mapping quality (mapq) filter of 11. We assessed a percent length coverage filter of a gene by reads of 10, 50, and 80% and chose the most permissive (10%) for comparison between the shotgun and enriched samples. These low thresholds were necessary for analyzing the shotgun data to obtain any reasonable results at all. Finally, we used a filter to check for the probes mapping to the reference sequences in most comparisons except to identify genes in the shotgun samples that would not be captured by our probeset. We repeated the same analysis process for the Negative Controls - blank
libraries. In Phase 2 Set 1, there were very few reads associated with the blank library after enrichment, so we used the raw sequencing reads for analysis. For the blank in Set 2, we omitted deduplication, and we could not identify any reads associated with the blank indexes after sequencing for Set 3. Read counts were normalized using the All Mapped Reads column in the gene mapping file and the reference length in kb along with the total number of reads available for mapping (per million) (RPKM). Hierarchical clustering was performed using Gene Cluster 3.0 and Java Tree View v 1.1.6r4 (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm) using a log transformation and clustering arrays with an uncentered correlation (Pearson) and average linkage. For rarefaction analysis, we first aligned trimmed reads against CARD (version 3.0.0) using Bowtie2, followed by filtering for mapping quality ≥ 11 (47). This file along with an annotation file for CARD was analyzed with the AmrPlusPlus Rarefaction Analyzer (http://megares.meglab.org/amrplusplus) (58) with subsampling every 1% of total reads and a gene read length coverage of at least 10%. The average number of genes identified after rarefaction was plotted and fit to a logarithmic curve to allow for simplified extrapolation. We generated heat maps and figures in Prism 8 for macOS (https://www.graphpad.com).

Data Access

Raw sequencing reads (FASTQ) for IIDR Clinical Isolate Collection bacterial isolate genome assembly were deposited in NCBI BioProject PRJNA532924. All metagenomic sequencing results, enriched or shotgun, were deposited in NCBI BioProject PRJNA540073. The probeset sequences and annotations are available at https://card.mcmaster.ca/download.

Acknowledgements

This research was funded by the Canadian Institutes of Health Research (MT-14981 to G.D.W., PJT-156214 to A.G.M.) and Genome Canada (to A.G.M.), by a Canada Research Chair (to G.D.W., H.N.P. and M.G.S.), and by a Cisco Research Chair in Bioinformatics supported by Cisco Systems Canada (to A.G.M.), Ontario Research Fund grant, and the Boris Family Foundation. A.K.G. was supported by a Canada Graduate Scholarship (Master’s) from the Natural Sciences and Engineering Council of Canada, an Ontario Graduate Scholarship, and an Ontario Graduate Fellowship. Computer
resources were supplied by the McMaster Service Lab and Repository computing cluster, funded in part by grants to A.G.M. from the Canadian Foundation for Innovation (34531). We would like to acknowledge Linda Ejim for clinical isolate culturing, Susan McCusker from the Center for Microbial Chemical Biology at McMaster University for genomic DNA isolation from clinical isolates, Laura Rossi for DNA isolation from human stool, and the McMaster Genomics Facility for next-generation sequencing. Additional thanks to Kara Tsang and Biren Dave for genome assembly and annotation of the clinical isolates, Nicholas Waglechner for helpful discussions and guidance pertaining to analyses, Christy Groves for valuable input on the figures, and all the members of the McMaster Ancient DNA Center.

Competing Interests

The authors declare no conflicts of interest.

1. Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, Vlieghe E, Hara GL, Gould IM, Goossens H, Greko C, So AD, Bigdeli M, Tomson G, Woodhouse W, Ombaka E, Peralta AQ, Qamar FN, Mir F, Kariuki S, Bhutta ZA, Coates A, Bergstrom R, Wright GD, Brown ED, Cars O. 2013. Antibiotic resistance—the need for global solutions. Lancet Infect Dis 13:1057–1098.

2. Levy SB, Marshall B. 2004. Antibacterial resistance worldwide: causes, challenges and responses. Nat Med 10:S122–S129.

3. Schwartz KL, Morris SK. 2018. Travel and the Spread of Drug-Resistant Bacteria. Curr Infect Dis Rep 20:29.

4. Gaze WH, Krone SM, Larsson DGJ, Li X-Z, Robinson JA, Simonet P, Smalla K, Timinouni M, Topp E, Wellington EM, Wright GD, Zhu Y-G. 2013. Influence of Humans on Evolution and Mobilization of Environmental Antibiotic Resistome. Emerg Infect Dis 19.

5. Surette MD, Wright GD. 2017. Lessons from the environmental antibiotic resistome. Annu Rev Microbiol 71:309–29.

6. Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J, Handelsman J. 2010. Call of the wild: Antibiotic resistance genes in natural environments. Nat Rev Microbiol 8:251–259.

7. Fitzpatrick D, Walsh F. 2016. Antibiotic resistance genes across a wide variety of metagenomes. FEMS Microbiol Ecol 92:1–8.

8. Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MOA, Dantas G. 2012. The shared antibiotic resistome of soil bacteria and human pathogens. Science (80-) 337:1107–1111.

9. Luo Y, Yang F, Mathieu J, Mao D, Wang Q, Alvarez PJJ. 2013. Proliferation of Multidrug-Resistant New Delhi Metallo-β-lactamase Genes in Municipal Wastewater Treatment Plants in Northern China. Environ Sci Technol Lett 1:26–30.

10. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. 2016. The structure and diversity
of human, animal and environmental resistomes. Microbiome 4:1–15.

11. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN, Doshi J, 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.

12. Walsh F, Duffy B. 2013. The Culturable Soil Antibiotic Resistome: A Community of Multi-Drug Resistant Bacteria. PLoS One 8:e65567.

13. Mezger A, Gullberg E, Göransson J, Zorzet A, Herthnek D, Tano E, Nilsson M, Andersson DI. 2015. A general method for rapid determination of antibiotic susceptibility and species in bacterial infections. J Clin Microbiol 53:425–432.

14. Zumla A, Al-Tawfiq JA, Enne VI, Kidd M, Drosten C, Breuer J, Muller MA, Hui D, Mauerer M, Bates M, Mwaba P, Al-Hakeem AA, Memish ZA, Gant V. 2014. Rapid point of care diagnostic tests for viral and bacterial respiratory tract infections-needs, advances, and future prospects. Lancet Infect Dis 14:1123–1135.

15. Pulido MR, García-Quintanilla M, Martín-Peña R, Cisneros JM, McConnell MJ. 2013. Progress on the development of rapid methods for antimicrobial susceptibility testing. J Antimicrob Chemother 68:2710–2717.

16. Crofts TS, Gasparrini AJ, Dantas G. 2017. Next-generation approaches to understand and combat the antibiotic resistome. Nat Rev Microbiol 15:422–434.
Poinar HN. 2016. 17th Century Variola Virus Reveals the Recent History of Smallpox. Curr Biol 26:3407–3412.

Devault AM, Mortimer TD, Kitchen A, Kiesewetter H, Enk JM, Golding GB, Southon J, Kuch M, Duggan AT, Aylward W, Gardner SN, Allen JE, King AM, Wright G, Kuroda M, Kato K, Briggs DEG, Fornaciari G, Holmes EC, Poinar HN, Pepperell CS. 2017. A molecular portrait of maternal sepsis from Byzantine Troy. Elife 6:1–31.

Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard J-M, Poinar HN. 2014. Ancient Whole Genome Enrichment Using Baits Built from Modern DNA. Mol Biol Evol 31:1292–1294.

Depledge DP, Palser AL, Watson SJ, Lai IY-C, Gray ER, Grant P, Kanda RK, Leproust E, Kellam P, Breuer J. 2011. Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. PLoS One 6:e27805.

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.

Noyes NR, Weinroth ME, Parker JK, Dean CJ, Lakin SM, Raymond RA, Rovira P, Doster E, Abdo Z, Martin JN, Jones KL, Ruiz J, Boucher CA, Belk KE, Morley PS. 2017. Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. Microbiome 5:142.

Allicock OM, Guo C, Uhlemann A-C, Whittier S, Chauhan L V., Garcia J, Price A, Morse SS, Mishra N, Briese T, Lipkin WI. 2018. BacCapSeq: a Platform for Diagnosis and Characterization of Bacterial Infections. MBio 9.

Mercer TR, Clark MB, Crawford J, Brunck ME, Gerhardt DJ, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. 2014. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat Protoc 9:989–1009.

Metsky HC, Siddle KJ, Gladden-Young A, Qu J, Yang DK, Brehio P, Goldfarb A, Piantadosi A, Wohl S, Carter A, Lin AE, Barnes KG, Tully DC, Corleiss B, Hennigan S, Barbosa-Lima G, Vieira YR, Paul LM, Tan AL, Garcia KF, Parham LA, Odia I, Eromon P, Folarin OA, Goba A, Simon-Lorière E, Hensley L, Balmaseda A, Harris E, Kwon DS, Allen TM, Runstadler JA, Smole S, Bozza FA, Souza TML, Isern S, Michael SF, Lorenzana I, Gehrk E, Bosch I, Ebeg G, Grant DS, Happi CT, Park DJ, Gnirke A, Sabeti PC, Matranga CB. 2019. Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. Nat Biotechnol 37:160–168.

Phillippy AM, Deng X, Zhang W, Salzberg SL. 2009. Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics 10:293.

Chafin TK, Douglas MR, Douglas ME. 2018. MrBait: Universal identification and design of targeted-enrichment capture probes. Bioinformatics 34:4293–4296.

Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Eukirchen G, Butte AJ, Snyder M. 2011. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29:908–914.

Shearer AE, Hildebrand MS, Ravi H, Joshi S, Guiffre AC, Novak B, Happe S, LeProust EM, Smith RJH. 2012. Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment. BMC Genomics 13:618.

Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. 2019. Contamination in Low Microbial Biomass Microbiome Studies: Issues and
Recommendations. Trends Microbiol 27:105–117.

37. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:1741–7007.

38. Sandalli C, Buruk CK, Sancaktar M, Ozgumus OB. 2010. Prevalence of integrons and a new dfrA17 variant in Gram-negative bacilli which cause community-acquired infections. Microbiol Immunol 54:164–9.

39. Wally N, Schneider M, Thannesberger J, Kastner MT, Bakonyi T, Indik S, Rattei T, Bedarf J, Hildebrand F, Law J, Jovel J, Steininger C. 2019. Plasmid DNA contaminant in molecular reagents. Sci Rep 9:1652.

40. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, Peacock SJ, Smith GCS, Parkhill J. 2018. Recognizing the reagent microbiome. Nat Microbiol 3:851–853.

41. Chiu CY, Miller SA. 2019. Clinical metagenomics. Nat Rev Genet 20:341–355.

42. Waseem H, Jameel S, Sales Ur Rehman H, Tauseef I, Farooq U, Jamal A, Ali MI. 2019. Contributions and Challenges of High Throughput qPCR for Determining Antimicrobial Resistance in the Environment: A Critical Review. Molecules 24:163.

43. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111–118.

44. Altschup SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. J Mol Biol 215:403–410.

45. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2017. GenBank. Nucleic Acids Res 45:D37–D42.

46. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

47. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Project G, Subgroup DP. 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma Appl NOTE 25:2078–2079.

49. Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

50. Rouillard JM, Zuker M, Gulati E. 2003. OligoArray 2.0: Design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res 31:3057–3062.

51. Whelan FJ, Verschoor CP, Stearns JC, Rossi L, Luinstra K, Loeb M, Smieja M, Johnstone J, Surette MG, Bowdish DME. 2014. The loss of topography in the microbial communities of the upper respiratory tract in the elderly. Ann Am Thorac Soc 11:513–521.

52. Bolger AM, Lohse M, Usadel B. 2014. Trimmmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

53. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.
29 

54. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. 2013. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics 29:2253–2260.

55. Meyer M, Kircher M. 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb Protoc 2010: pdb.prot5448-pdb.prot5448.

56. Jiang H, Lei R, Ding S-W, Zhu S. 2014. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15:182.

57. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. Bamtools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics 27:1691–1692.

58. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, Rovira P, Abdo Z, Jones KL, Ruiz J, Belk KE, Morley PS, Boucher C. 2017. MEGARes: An antimicrobial resistance database for high throughput sequencing. Nucleic Acids Res 45:D574–D580.
Figure 1 – A platform for capture and identification of diverse antibiotic resistance genes. The targeted capture sequencing workflow begins with DNA isolation from a sample of interest (stool from a healthy donor in this example). a) DNA is fragmented through sonication and prepared as a sequencing library. b, c) Target sequences representing less than 1% of the total DNA are captured through hybridization with biotinylated probes and streptavidin-coated magnetic beads. d, e) The captured and amplified library fragments are sequenced, and reads are analysed for AMR sequence content by mapping to CARD.

Figure 2 – Design of a probeset to target over 2000 antibiotic resistance genes. Breakdown of resistance gene classes from CARD that are targeted by probes. A legend for the top 10 classes is shown; AME = aminoglycoside modifying enzymes, qnr = quinolone resistance genes. The remaining 122 genes belong to various classes. The beta-lactamase genes make up the majority of genes targeted by probes and are highlighted with a black border.

Figure 3: Comparison of enriched and shotgun results for on-target mapping, recovery, and length coverage. Each point on the graph represents a replicate experiment either as a genome that was enriched individually or when pooled with other genomes across both trials. The horizontal line for each isolate represents the mean. A) The percentage of reads on-target for each bacterium tested in various sample types for both enriched and shotgun samples. B) The percent recovery of regions predicted to be targeted by probes for each bacterial genome tested in both enriched and shotgun samples (1 versus 10 versus 100 reads per probe-targeted region). C) The average percent length coverage of probe-targeted regions with reads from isolates tested individually and in pools in both enriched and shotgun samples (1 versus 10 versus 100 reads). Samples were excluded from panel C if they did not have any probe-targeted regions with a given read coverage. This represents 6 samples in the “At least 10 reads” panel (all from the shotgun data (C0002 – 1; C0050 -2; C0060 – 2; C006 -1; C0292 – 1) and all samples for the shotgun data in the “At least 100 reads panel” and 5 for the enriched (C0060 – 4; C0292 – 1).

Figure 4: Enrichment results in higher read counts on antibiotic resistance genes compared to shotgun sequencing. Normalized read counts at each probe-targeted region within the Escherichia coli C0002 genome (A) and Staphylococcus aureus C0018 genome (B) in enriched and shotgun samples including individual and “mock metagenomes” of multiple isolates. Among enriched and shotgun pairs, reads were subsampled to equal depths and mapped to the individual isolate’s genome. Read counts were normalized by number of reads mapping per target length in kilobases per million reads. The predicted number of probes for each region along the genome are shown in the panels below. The Y axes are in the logarithmic scale.

Figure 5: Comparison of resistance elements between enriched and shotgun libraries. For the enriched and shotgun samples, the full number of reads for each sample were mapped to CARD using rgi bwt and the results were filtered for genes with probes mapping, with reads with an average mapping quality >=11 and a percent length coverage of a gene by reads greater than or equal to 10%. A) i) Read counts were normalized per kilobase of reference gene per million reads sequenced (RPKM) and log transformed to produce the heatmap. The rows are grouped based on resistance mechanisms as annotated in CARD (not all mechanisms and classes are labelled). ABC = ATP-binding cassette antibiotic efflux pump; MFS = major facilitator superfamily antibiotic efflux pump; RND = resistance-nodulation cell division antibiotic efflux pump; MLS = macrolides, lincosamides, streptogramins. ii) The number of reads used for mapping in each sample. B) Left - Overlap of genes found with at least 10 reads, a percent coverage greater than or equal to 10% and an average mapping quality of reads greater than or equal to 11 in the 27 enriched and 6 shotgun samples. Between all samples, enriched or shotgun sequenced, there were 89 genes with reads passing these filters; 13 overlap, 57 are unique to the enriched, and 19 are...
unique to the shotgun samples. Right - Of the 19 genes only identified through shotgun sequencing, only
4 of these genes are predicted to be targeted by probes.

Table 1: Comparing genes with reads for shotgun and enriched stool library pairs. We mapped the
full number of reads from shotgun and enriched pairs to CARD using rgi bwt. Results samples were
filtered for genes with at least 10 reads, those probes mapping (only for the enriched samples), average
read mapping quality >=11 and average read length coverage >=10%. Filtered genes and their normalized
read counts (RPM) from each enriched/shotgun pair were combined to compare and determine the fold-
enrichment.

| Probes (ng) | Library (ng) | Fold-difference in reads (enriched vs shotgun) | Genes found in shotgun | Genes found in enriched | Genes overlapping | Genes with probes missing in enriched | Fold-enrichment (min – max) |
|-------------|--------------|-----------------------------------------------|------------------------|------------------------|-------------------|--------------------------------------|-----------------------------|
| Set 1       |              |                                               |                        |                        |                   |                                      |                             |
| 200         | 100          | 389.70                                        | 18                     | 49                     | 9                 | 1                                    | 1054.92 (0 – 10905.8)       |
| 100         | 200          | 82.24                                         | 20                     | 25                     | 7                 | 5                                    | 879.87 (0 – 6459.8)         |
| Set 2       |              |                                               |                        |                        |                   |                                      |                             |
| 400         | 200          | 154.93                                        | 27                     | 55                     | 12                | 4                                    | 868.16 (0 – 9612.1)         |
| 100         | 100          | 80.73                                         | 23                     | 61                     | 11                | 1                                    | 732.16 (0 – 8193.3)         |
| Set 3       |              |                                               |                        |                        |                   |                                      |                             |
| 100         | 100          | 66.67                                         | 19                     | 57                     | 9                 | 2                                    | 690.19 (0 – 6962.7)         |
| 25          | 50           | 88.26                                         | 22                     | 58                     | 9                 | 2                                    | 690.19 (0 – 7319.6)         |
Figure 5

A

Mechanism Class

Antibiotic efflux

ABC

MFS

RND

Antibiotic Inactivation

Aminoglycoside

beta-lactamase

Other

MLS

Target Alteration

Glycopeptide

Peptide

Tetracycline

Target Protection

Replacement

Enriched Shotgun

Read counts

174,396 0

Figure 5

Paired reads

B

Genes with >= 10 reads

n = 89

Genes not found in enriched

57

13

19

4

Enriched (27)

Shotgun (6)

With probes

Without probes

Figure 5
Figure 4
Figure 3
Total = 2021
Class A beta-lactamase (660)
Class D beta-lactamase (294)
Class C beta-lactamase (246)
Efflux systems (179)
Class B beta-lactamase (158)
AMEs (134)
qnr (95)
Glycopeptide resistance (67)
23S rRNA methyltransferase (37)
Trimethoprim resistance (29)
Other (122)

Figure 2:
Figure 1: