Context-Seq: CRISPR-Cas9 Targeted Nanopore Sequencing for Transmission Dynamics of Antimicrobial Resistance

Erica R. Fuhrmeister[1,2], Sooyeol Kim[2], Shruteek A. Mairal[3], Caroline McCormack[2], Benard Chieng[4], Jenna M. Swarthout[5], Abigail Harvey Paulos[2], Sammy M. Njenga[4], Amy J. Pickering[2,6*]

[1]Department of Environmental and Occupational Health Sciences, School of Public Health, University of Washington, Seattle, Washington, USA

[2]Department of Civil and Environmental Engineering, University of California, Berkeley, CA, USA

[3]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA, USA

[4]Eastern and Southern Africa Centre of International Parasite Control, Kenya Medical Research Institute, Nairobi, Kenya

[5]Department of Civil and Environmental Engineering, Tufts University, Medford, MA, USA

[6]Blum Center for Developing Economies, University of California, Berkeley, CA, USA

* Correspondence to: Dr. Amy Pickering, Civil and Environmental Engineering, University of California, Berkeley, CA, USA, 94720, pickering@berkeley.edu

Abstract: 209

Word count: 4946

Figures: 4

Abstract

Antimicrobial resistance (AMR) aligns with a One Health framework in that resistant bacteria and antibiotic resistance genes (ARGs) can be transmitted between humans, animals, and the environment. However, there is a critical need to more precisely understand how and to what extent AMR is exchanged between animals and humans. Metagenomic sequencing has low detection for rare targets such as ARGs, while whole genome sequencing of isolates is burdensome and misses exchange between uncultured bacterial species. We developed a novel, targeted sequencing assay using CRISPR-Cas9 to selectively sequence ARGs and their genomic context with long-read sequencing. Using this method, termed Context-Seq, we investigated overlapping AMR elements containing the ARGs *bla*$_{\text{CTX-M}}$ and *bla*$_{\text{TEM}}$ between adults, children, poultry, and dogs in animal-owning households in Nairobi, Kenya. We identified 22 genetically distinct clusters (> 80%ID over ≥ 3000 bp) containing *bla*$_{\text{TEM}}$ and one cluster containing *bla*$_{\text{CTX-M}}$ that were shared within and between households. Half of the clusters were shared between humans and animals, while the other half were shared only between animals (poultry-poultry, dog-dog, and dog-poultry). We identified potentially pathogenic hosts of ARGs including *Escherichia coli, Klebsiella pneumonia*, and *Haemophilus influenzae* across sample types. Context-Seq complements conventional methods to obtain an additional view of bacterial and mammalian hosts in the proliferation of AMR.

Introduction

Antimicrobial resistance (AMR) is a global challenge that threatens to undermine modern medicine. It is estimated that in 2019, nearly 5 million deaths were associated with bacterial antimicrobial resistance.[1] The global burden of AMR disproportionately falls on low-income countries, where high rates of illness, unregulated antibiotic usage, and limited access to sanitation infrastructure contribute to the selection and spread of AMR bacteria.[2,3] Current approaches to control antibiotic resistance rely on antibiotic stewardship; however, this approach is difficult in low-income countries where unregulated antibiotic usage in humans and animals is common and the burden of infectious diseases is high. Between 2000 and 2015, antibiotic drug consumption rates increased by 77% in low- and middle-income countries, compared to a decrease of 4% in high income countries.[4]

AMR can be shared between humans, animals, and the environment. For this reason, AMR fits within the One Health framework, which integrates human, animal, and environmental health to tackle complex public health problems. AMR can spread through the dissemination of whole bacteria carrying resistance, as well as through horizontal gene transfer of mobile elements, including between benign and pathogenic bacteria. In addition, soil, water, and air are known environmental reservoirs of AMR.[5–8] From a molecular perspective, determining the genomic context of ARGs is critical for studying the proliferation of AMR as it can allow for the identification of mobile elements, co-occurring genes, and host bacteria.[9] These additional pieces of information may yield insights on the mechanisms of exchange between reservoirs and the role of zoonotic transmission. In order to curb the spread of AMR, we need to be able to identify the most important transmission pathways (e.g., poultry, dogs, water, soil) in a given setting.

Current methods to investigate overlap in AMR elements between reservoirs primarily rely on culturing and whole genome sequencing of isolates.[10–12] However, culturing only captures a small fraction of organisms, with a bias towards bacteria more fit for selective conditions. Culture-independent methods, such as metagenomic sequencing, require high sequencing depth to capture low-abundance targets such as ARGs.[13] In addition, untargeted metagenomic sequencing can waste millions of reads per sample with very low coverage of medically important ARGs. This abundance of data can be costly to store and require significant computing resources to analyze.

Targeted sequencing approaches are promising for studying AMR by enrichment of genomic regions of interest. Illumina probe capture is inherently limited by the read length (200-500 bps), which makes investigating ARGs in their genomic context infeasible. Recently, Cas9-based enrichment has been applied to short and long read sequencing for clinical applications.[14,15] In brief, extracted DNA is dephosphorylated followed by Cas9 cutting facilitated by guideRNAs. Sequencing adapters are selectively ligated to only the d(A) tailed ends that result from Cas9 cutting. For example, Cas9-guided adapter ligation was used to perform multiplexed detection of ARGs in human blood spots with Illumina short-read sequencing[14] and to investigate human alleles in breast tissue with Oxford Nanopore long reads.[15] The selectivity introduced through guideRNAs coupled with long-read sequencing that can capture long DNA fragments make this a promising approach to investigate ARGs within their genomic context.

In this study, we developed and optimized a Cas9 targeted sequencing assay to selectively sequence ARGs and their genomic context, hereby referred to as Context-Seq. We demonstrate the utility of Context-Seq by applying the method to detect the ARGs *bla*$_{\text{CTX-M}}$ and *bla*$_{\text{TEM}}$ in human (adult and child), poultry, and canine fecal samples collected from households in Nairobi, Kenya to investigate overlapping antimicrobial resistance elements.

Results

Samples and ARG Target Selection. In order to select relevant ARG targets for Context-Seq, we first analyzed available human and animal fecal samples collected from seven households in Nairobi, Kenya[16] using a Taqman Array Card to detect 14 ARGs and eight pathogen targets (**Table S1** and **Figure S1**). All samples were positive for *tetA*, *sul1*, and *bla*$_{\text{TEM}}$. *Bla*$_{\text{NDM}}$ and *mcr-1* were detected in canine and poultry samples only (*mcr-1*: 75% canine, 7% poultry; *bla*$_{\text{NDM}}$: 50% canine, 13% poultry). *Bla*$_{\text{CTX-M Group 1}}$, *bla*$_{\text{CTX-M Group 9}}$, *bla*$_{\text{OXA-10}}$, and *bla*$_{\text{SHV}}$ were detected in at least 50% of samples of each type. We selected two clinically relevant targets in high abundance (*bla*$_{\text{TEM}}$ and *bla*$_{\text{CTX-M}}$) and samples from four households that were positive for *bla*$_{\text{TEM}}$ and *bla*$_{\text{CTX-M group 1}}$ (**Table S2**).

While many metagenomic approaches capture hundreds to thousands of resistance genes, not all resistance genes are clinically important.[17] Extended-spectrum beta-lactamases (ESBLs) genes are of

high medical importance as they can confer resistance to most beta-lactams including cephalosporins.[18] *Bla*$_{CTX-M}$ is a globally distributed gene group where all alleles are considered ESBLs.[19] Common genotypes include *bla*$_{CTX-M-15}$ and *bla*$_{CTX-M-14}$.[19] *Bla*$_{CTX-M}$ alleles have been found in humans,[20] animals,[21] wastewater,[22] and other environmental reservoirs.[23] They are also present in clinical isolates of *Escherichia coli*, *Klebsiella pneumoniae*, *Salmonella* species, *Pseudomonas aeruginosa*, and other bacterial taxa.[24] Similarly, *bla*$_{TEM}$ are globally distributed, present in multiple reservoirs, and found in clinical isolates.[21,25–28] However, *bla*$_{TEM}$ alleles differ in phenotypic resistance conferred, ranging from penicillin resistance (e.g., *bla*$_{TEM-1}$) to ESBLs (e.g., *bla*$_{TEM-10}$).[29]

Guide Design Tool. To design Cas9 guide RNAs, we utilized existing software (CHOPCHOP)[30] and developed a custom script to estimate off-target activity of guides in complex microbial communities, which is publicly available (https://github.com/Shruteek/Optimized-sgRNA-Design). Guides were selected based on high CHOPCHOP predicted efficiency, genomic location near the ends of the genes, and lower predicted off-target activity (**Table S3**). Notably, predicted efficiency as well as off-target activity are based on empirical data[31] and may not be well representative of real systems. When comparing guides for *bla*$_{TEM}$, all pairs resulted in enrichment; there was, however, variation in enrichment based on the combined pair (**Figure S2**) and best performing guides were not necessarily predicted to have the highest on target activity and lowest off-target activity (**Table S3**). To facilitate capture of genomic context in both directions of a target ARG, different fractions of the sample DNA were cut by guides on the sense and antisense strands separately and then pooled. Alleles targeted by guides are indicated in the supporting information (**Tables S4-S5**).

Protocol Optimization. We optimized long-read Cas9 enrichment, originally validated for variant detection in cell culture and human tissue samples,[15] to detect ARGs in fecal and soil samples. We investigated these modifications on a mock community comprised of an *Escherichia coli* isolate with *bla*$_{CTX-M-55}$ and *bla*$_{TEM-1}$ genes spiked into a composite sample of extracted DNA from Kenyan soil. Modifications evaluated included adaptive sequencing, multiple guides per target per strand, longer

incubation time for Cas9 cleavage, and the addition of thermolabile Proteinase K. We also evaluated the impact of including two targets in the model system and in a human fecal sample.

Our final protocol enriched for two targets (*bla*$_{CTX-M}$ and *bla*$_{TEM}$) in two directions and added a thermolabile Proteinase K digestion to previously published methods (**Figure 1A**). Adaptive sequencing, longer Cas9 digestion, and additional guides per target did not improve the assay performance (**Figure 1B**) but utilizing thermolabile Proteinase K after the Cas9 digestion did (**Figure 1C**). Inclusion of guides for both *bla*$_{CTX-M}$ and *bla*$_{TEM}$ decreased enrichment for both targets in our mock system (**Figure 1B**) and in the human fecal sample (**Figure 1D and 1E**). Since the decrease in enrichment was minor for *bla*$_{TEM}$ (non-normalized coverage [mean (std)]: two targets [1265 (253)]; one target [1517 (337)]), the target in higher abundance, we processed samples enriching for both targets. All protocol modifications resulted in an enrichment of 7-15X coverage over untargeted methods (**Figure 1B**).

Using our final protocol, we sequenced 13 fecal samples across four households (five human, three canine, and five poultry) on individual MinION flow cells (**Table S6**). The percentage of reads that aligned to either *bla*$_{CTX-M}$ and/or *bla*$_{TEM}$ out of the total reads that passed quality filtering was 0.4% (range 0.02-2.01%). Of the reads that aligned to *bla*$_{TEM}$ across all samples, the average length was 4854 [std 1081] base pairs (bps) and of those that aligned to *bla*$_{CTX-M}$ the average length was 4381 [745] bps. Clustering reads (>1500 bps) containing ARGs at 85% identity resulted in an average of 39 [35] clusters per sample (**Table S7**).

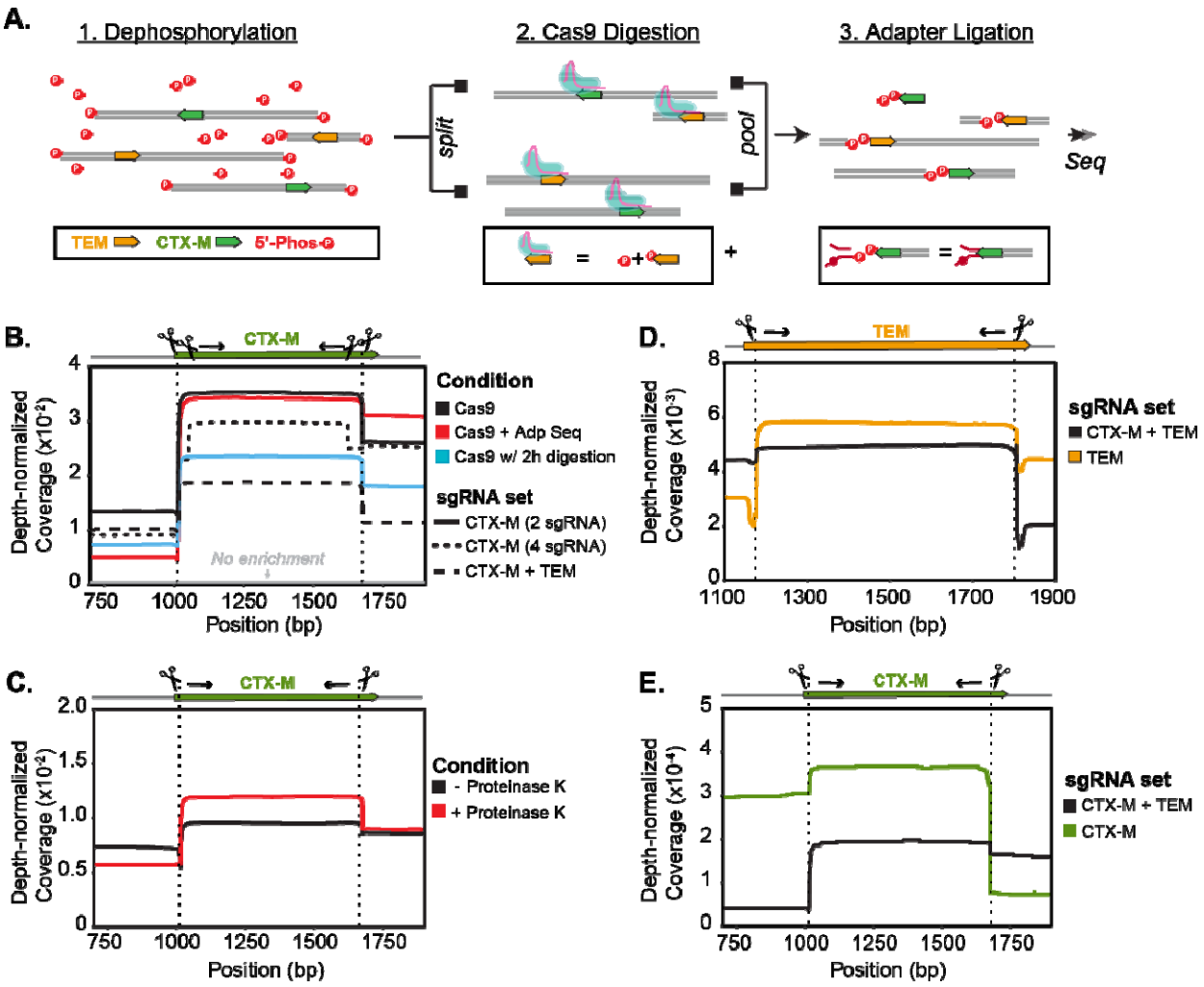

Figure 1. Cas9 is used to selectively sequence DNA fragments containing *bla*$_{CTX-M}$ and *bla*$_{TEM}$. A) Schematic of Context-Seq workflow involving dephosphorylation, library splitting for Cas9 cutting on each strand, and adapter ligation. Phosphorylated ends are indicated with a red P. **B)** Comparison between library preparation modifications (adaptive sequencing, longer Cas9 digestion, and additional guides) against no enrichment in a mock system. Depth normalized coverage is calculated by dividing coverage by the total reads obtained per each sequencing run. **C)** Comparison between conventional Cas9 enrichment protocol and the inclusion of Proteinase K following Cas9 digestion in a mock system. **D)** Normalized coverage of *bla*$_{TEM}$ in a human fecal sample comparing enrichment of TEM alone (yellow) and both TEM and CTX-M (black). **E)** Normalized coverage of CTX-M in a human fecal sample comparing enrichment of CTX-M alone (green) and both TEM and CTX-M (black).

Context-Seq enabled identification of ARGs and annotation of their surrounding genomic context. Annotated sequences resulted in the expected cut pattern based on guide design. Sequences begin with an ARG cut on either the sense or antisense strand and contain additional annotated genes across variable long read lengths (**Figure 2**). Sequences for *bla*$_{CTX-M}$ ranged from 1,662-17,369 bps and included mobile genes annotated as integration/excision (e.g., *tnpA, tnpR, hpaI, gin*),

replication/recombination/repair (e.g., *dnaQ, repL, impB*), and phage (e.g., *ant, gp23, kilA, orf16*) (**Figure 2A**). Sequences for *bla*$_{TEM}$ ranged from 1,489-23,336 bps and included mobile genes annotated as integration/excision (e.g., *tnpA, tnpR, int*, IS6 family transposases), transfer (*mob, finO, traI*), replication/recombination/repair (e.g., *rop, repC, repM, parM*), and phage (*bof, cre, pacB*) (**Figure 2B**). Co-occurring ARGs captured by Context-Seq included *aph(6)-Id*, *aph(3")-Ib, mphA, qnrS1, sul3,* and *bla*$_{CTX-M}$ among others. We also identified *bla*$_{TEM}$ co-occurring with disinfectant resistance (*qacEdelta1*) and mercuric reductase (*merA, merT, merC*) genes. Eight sequences containing *bla*$_{TEM}$ were greater than 18,000 bps and included multiple co-occurring ARGs. For example, in the consensus sequences ≈ 20 kbp in the canine sample, *bla*$_{TEM}$, *sul2*, *aph(3")-Ib*, *aph(6)-Id*, *mrx*, *mphA* were identified across the 23,336 bp sequence. In the consensus sequence for the adult human fecal sample in household two, *bla*$_{TEM}$, *dfrA8*, *sul2*, *aph(3")-Ib*, *aph(6)-Id* occurred across the 19,716 bp sequence (**Figure 2C**).

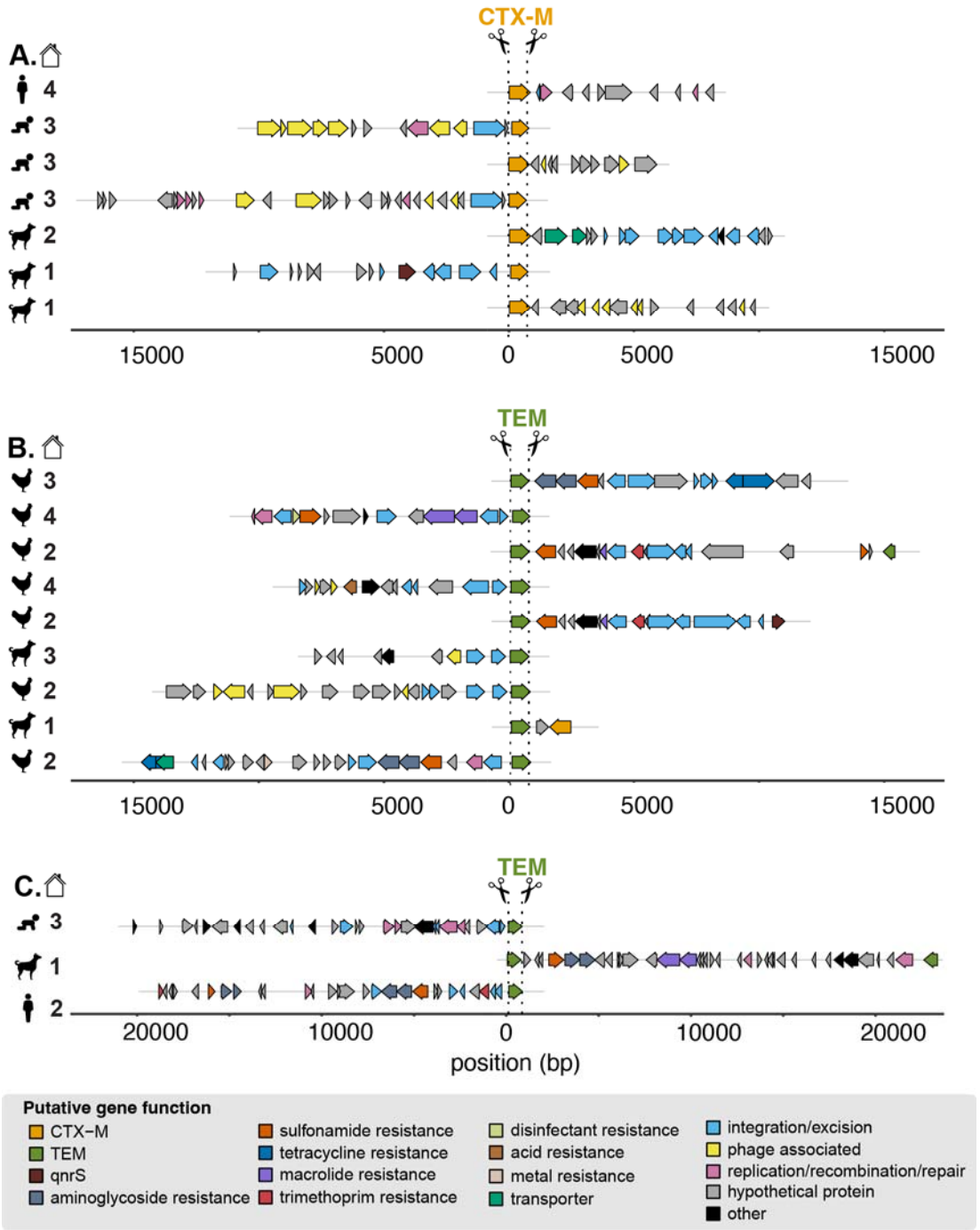

Figure 2. Annotated enriched sequences containing *bla*$_{CTX-M}$ and *bla*$_{TEM}$ generated by Context-Seq. A subset of sequences containing **A)** *bla*$_{CTX-M,}$ **B)** *bla*$_{TEM}$, and **C)** *bla*$_{TEM}$ ≈ 20,000 bp annotated for ARGs and mobile genetic elements. Sample type is indicated by symbol (child, adult, poultry, canine) and household by number. Note these sequences are a subset and some consensuses sequences were obtained from the same sample. In addition, all ARGs were aligned in the same orientation representing how cutting can proceed from the sense or anti-sense strand.

ARGs were identified on plasmids and primarily in E. coli, K. pneumoniae, and Hemophilus influenzae. Of the taxonomic identifications that were identified via Kraken2 and confirmed via BLASTn,

195 *bla*$_{TEM}$ was identified on plasmids and chromosomes (consensus sequences not identified as plasmids
196 with a plasX confidence <0.5) annotated as *K. pneumoniae*, *H. parainfluenzae*, *H. influenzae*, *E. coli* and
197 Enterobacteriaceae (**Figure 3A**). *Bla*$_{CTX-M}$ was identified on plasmids and chromosomes in *K.*
198 *pneumoniae* and *E. coli* (**Figure 3B**). Across hosts and genes (*bla*$_{TEM}$ and *bla*$_{CTX-M}$), the majority of
199 sequences were identified as *E. coli* followed by *K. pneumoniae*. Other gammaproteobacteria, *H.*
200 *parainfluenzae* and *H. influenzae,* were only found in human stool samples and not in animals.
201 We compared our method (Context-Seq in total DNA extracts) to a parallel study that cultured *E.*
202 *coli* without antibiotics and sequenced up to 5 pooled isolates from the same samples.[16] We calculated
203 the average coverage of the assembled contigs that resulted from Illumina sequencing and assembly of
204 the cultured *E. coli* isolates. All instances of *bla*$_{TEM}$ and *bla*$_{CTX-M}$ identified in cultured *E. coli* by Illumina
205 sequencing were also identified by Context-Seq in the same sample. In human fecal samples, the median
206 coverage of the contigs was higher with Illumina sequencing compared to Context-Seq; coverage was
207 160 with Illumina sequencing (range: [17-735]) and 49 [7-1237] with Context-Seq (paired two-sided t-test
208 p=0.82). In animal samples, the median coverage with Context-Seq (171 [66-3852]) was greater than
209 Illumina (138[19-1099]), although not statistically significant (paired two-sided t-test p=0.24) (**Figure 3C**).
210 In one poultry and one canine sample, no contigs containing *bla*$_{TEM}$ or *bla*$_{CTX-M}$ were assembled from the
211 cultured isolates, but both ARGs were identified in the canine and *bla*$_{TEM}$ was identified in the poultry
212 using Context-Seq.

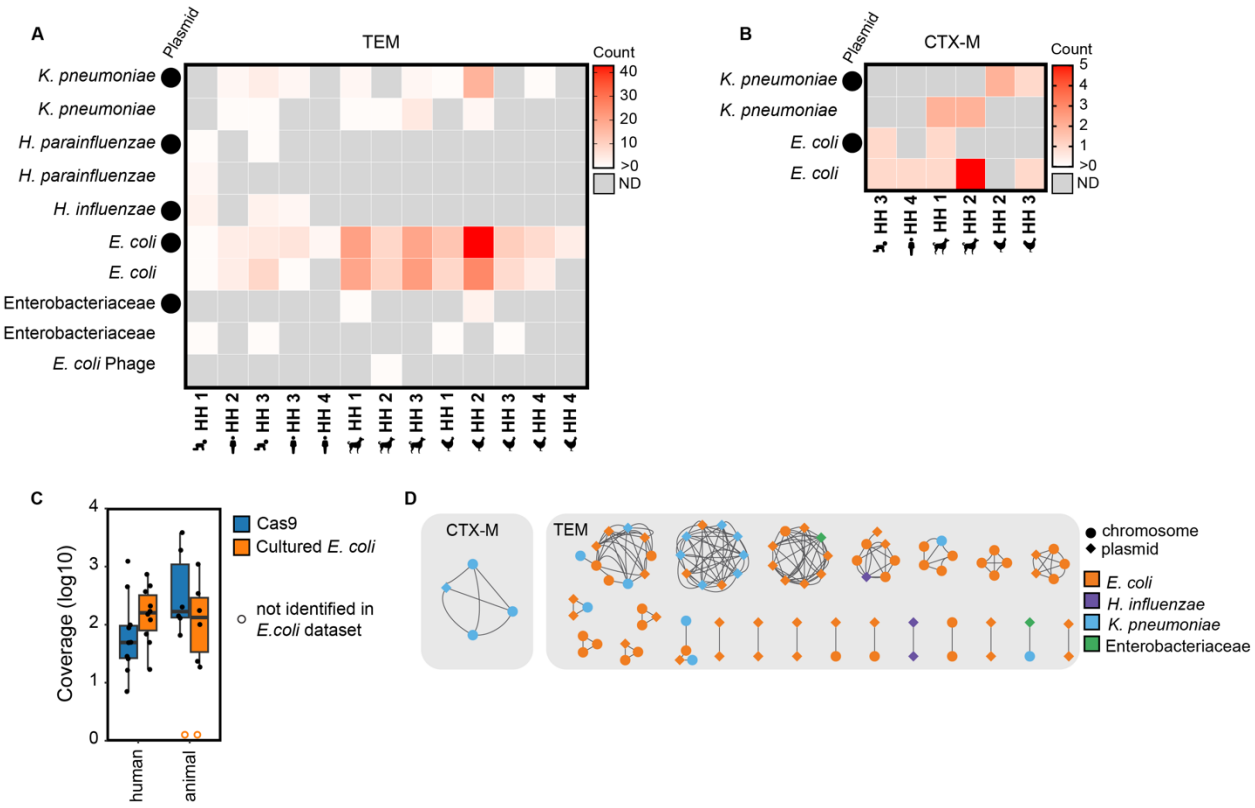

Figure 3. Taxonomic identifications of Context-Seq sequences containing *bla*$_{TEM}$ and *bla*$_{CTX-M}$. A) The number of sequences containing *bla*$_{TEM}$ identified as each taxa by Kraken2 and confirmed by BLASTn. Plasmids (plasX ≥ 0.5) are indicated by a circle. **B)** The number of sequences containing *bla*$_{CTX-M}$ identified as each taxa by Kraken2 and confirmed by BLASTn. Plasmids (plasX ≥ 0.5) are indicated by a circle. **C)** Comparison of Context-Seq to cultured and sequenced *E. coli* in the same sample. Coverage of the assembled contigs that resulted from 1.) the cultured *E. coli* using Illumina sequencing of the isolates and 2.) Context-Seq. Box shows interquartile range (25th to 75th percentiles) with the median and whiskers extending to 1.5 times the interquartile range. An open orange circle indicates no *bla*$_{TEM}$ or *bla*$_{CTX-M}$ was identified in that sample using Illumina. **D)** Taxonomic identification (color) and plasmid or chromosome designation (shape) in the 23 clusters (> 80%ID over ≥ 3000 bp) that are shared between hosts (human, poultry, canine) and households.

ARGs are shared between human and animal hosts and across households. A total of 23 clusters (> 80%ID over ≥ 3000 bp), one with *bla*$_{CTX-M}$ and 22 with *bla*$_{TEM}$, were shared between samples (**Figure 4A**). 11 clusters were shared between humans and animals, 11 shared between animals only, and one cluster was shared between humans only (**Figure 4B**). Of the animal-animal host sharing, seven clusters were observed in canines and poultry, one in canines only, and three in poultry only. 18 of the 23 clusters were found in more than one household and five were shared within individual households (**Figure 4B**). Within the shared clusters, the *bla*$_{CTX-M}$ gene aligned to a group of *bla*$_{CTX-M}$ alleles highly similar in the target region (CTX-M-15/224/238/163/194/232). The *bla*$_{TEM}$ genes in the shared clusters aligned primarily to two

groups of highly similar alleles (group 1: TEM-214/206/243/141/209/166 and group 2: TEM-217/234/104/198/228/135). Shared clusters generally contained multiple ARGs (e.g., *sul2*, *dfrA*, *tetA*, *aph(3")-Ib*) conferring resistance to sulfonamides, trimethoprim, tetracycline, and aminoglycosides in addition to the bla_{TEM} or bla_{CTX-M} targets. *TnpA,* which encodes for the transposase for transposon Tn3, was the most common integration/excision gene while *repA*, *repC*, and *parM* were the most common replication/recombination/repair genes.

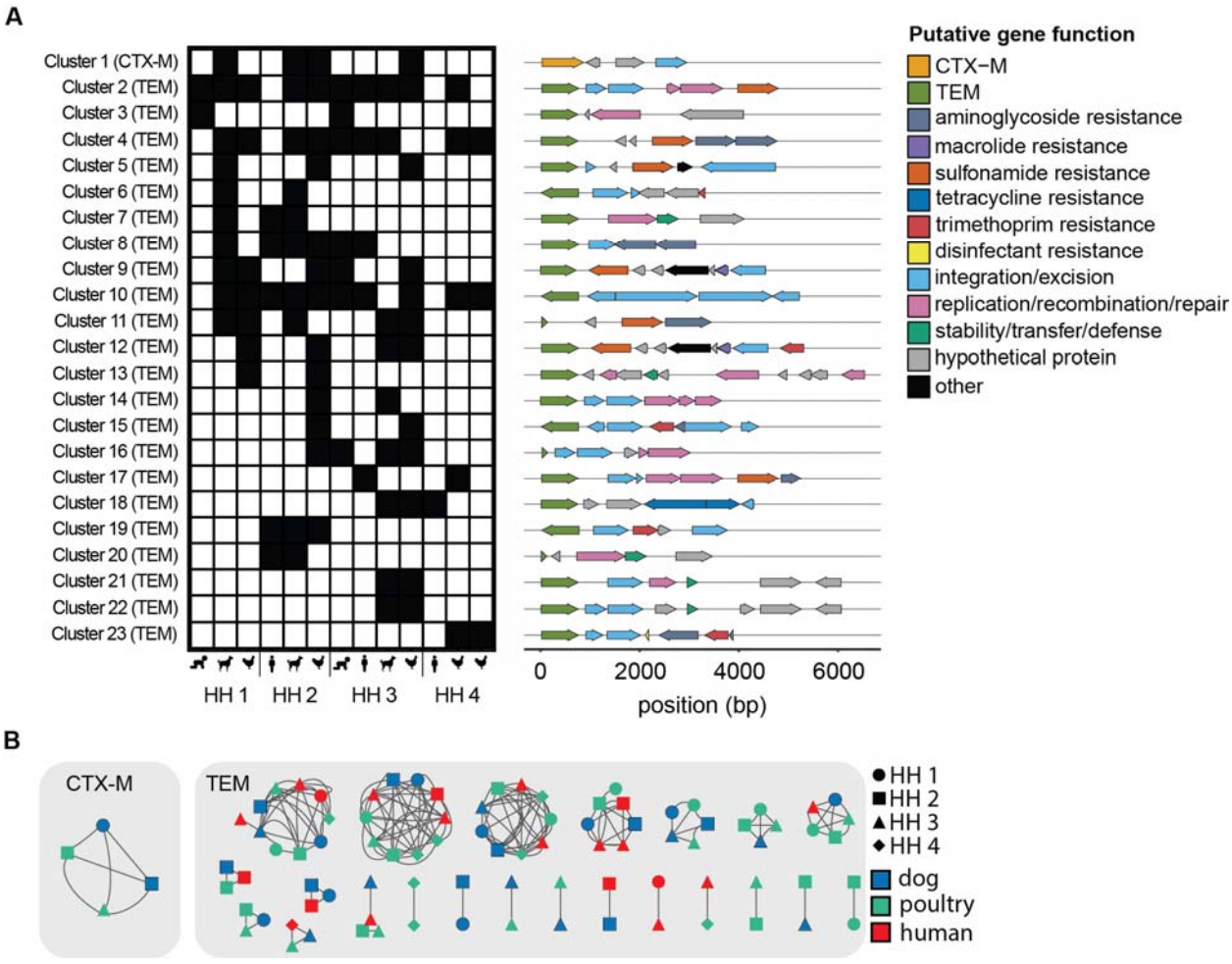

Figure 4. 23 clusters containing bla_{TEM} or bla_{CTX-M} were shared between samples. A) Presence (black)/absence of shared clusters (> 80%ID over ≥ 3000 bp) by sample type and household. ARG and mobile genetic element annotation of shared clusters. **B)** Shared clusters annotated by household (shape) and host type (color).

ARGs were shared across hosts through plasmids and chromosomes. The bla_{CTX-M} cluster was identified as *K. pneumoniae,* shared between three households, and shared between canines and poultry (**Figure 3D** and **Figure 4B**). The sequence containing the bla_{CTX-M} cluster was identified as a likely

plasmid in one of the four samples (**Figure 3D**). *Bla*$_{TEM}$ was shared across households and hosts on consensus sequences identified as *E. coli*, *K. pneumonia*, and *H. influenzae* (**Figure 3D**). Of the shared *bla*$_{TEM}$ clusters, approximately half were shared on the same element (plasmid to plasmid or chromosome to chromosome) and half were shared between elements (plasmids and chromosomes). The sole human to human shared cluster (cluster 3 between children in HH 1 and 3) was classified as a *H. influenza* plasmid in both samples.

Discussion

Here we developed a novel method, Context-Seq, using Cas9 targeted sequencing paired with long-read sequencing to enrich for fragments of DNA containing two clinically relevant ARGs, *bla*$_{TEM}$ and *bla*$_{CTX-M}$. We identified sharing of ARGs and genomic context between humans and animals, as well as between poultry and canines in urban Kenyan households, emphasizing the importance of the One Health approach to combatting AMR. Suspected hosts of ARGs were not just limited to *E. coli* but also included *K. pneumonia*, *H. influenzae*, and *H. parainfluenzae*. *E. coli* remains one of the most commonly characterized antimicrobial-resistant organism,[32] but many others are important to investigate in the context of AMR transmission.

An important finding of this work is the occurrence of non-*E. coli* hosts, primarily *K. pneumoniae* and *H. influenzae*. *K pneumoniae* is a leading cause of antimicrobial resistant infections,[33,34] especially in hospitalized patients[35] and is on the WHO global priority pathogens list.[36] While *K. pneumoniae* can survive in multiple environments including soil, water, and the intestinal tract of humans and animals,[35] few studies have investigated resistant *K. pneumoniae* from a molecular epidemiology, One Health perspective (i.e., strain sharing in humans, animals, and the environment). A previous study in Kenya collected *K. pneumoniae* from community fecal samples, healthcare-associated fecal samples, and hospital surfaces across multiple counties. Among extended spectrum beta-lactamase (ESBL) isolates, *bla*$_{CTX-M-15}$ and *bla*$_{TEM-181}$ were the most common genes.[37] We identified *bla*$_{CTX-M-15\ like}$ in *K. pneumoniae* in a shared cluster, while *bla*$_{tem-181\ like}$ was identified in one sequence, but not in a shared cluster. Most *bla*$_{TEM}$ genes were highly similar to non-ESBL genes *bla*$_{TEM-141}$ and *bla*$_{TEM-135}$ (groups 1 and 2 above). Very few studies have been conducted on *K. pneumoniae* in household animals, although one study investigated

K. pneumoniae isolated from raw meat samples in Nairobi.[38] They found high resistance to ampicillin but very few isolates had ESBL resistance.[38] Further, *H. influenzae* is a leading cause of bacterial respiratory infection[39] with recent increasing resistance to beta-lactams.[40] Compared to *K. pneumoniae* and *E. coli*, there are even fewer studies of resistant *H. influenzae,* especially in LMICs. One study investigated resistant *H. Influenzae* in Morocco and found one-third of isolates carried resistance genes to beta-lactams.[41] Most were susceptible to ESBLs, primarily demonstrating resistance to ampicillin and amoxicillin.[41]

Approximately half of shared ARGs and their genomic context (11/23 clusters) were shared between animals. An extensive previous study of *E. coli* across hosts in Nairobi (the UrbanZoo project) found highly similar resistomes among livestock poultry, both within and between households.[11] This work hypothesized the similar poultry resistomes were the result of similar antimicrobial selective pressure since use of antimicrobials for therapeutic or prophylactic purposes is consistent across Nairobi.[42] We observed significant overlap between poultry and canines (7 out of 11 animal-animal clusters), however the UrbanZoo project did not investigate canines. A similar mechanism may exist for canines and poultry in that they may consume similar antibiotics, and their guts could select for similar ARGs.[43] Another potential mechanism of AMR acquisition in canines is through scavenging. Previous work has demonstrated that the widespread waste (including human and animal feces as well as garbage) across the urban landscape of Nairobi can serve as a reservoir for AMR.[44] Canines could acquire similar AMR as poultry through eating of poultry feces.

Human and animal overlap was observed in the other half of shared clusters (11 out of 23). Previous evidence for human and animal resistome sharing has been mixed.[10,12,45,46] In a related study, where *E. coli* was isolated from humans, animals, and the environment in the same households as our work, human and animal strain sharing was rare.[16] The majority of *E. coli* strain similarity was observed between humans and stored drinking water, and poultry and soil, implicating the environment as a reservoir between hosts.[16] The observed higher relative degree of sharing between humans and animals in our work is likely due to method specific differences. Here, we investigated ARG containing DNA fragments only and our metagenomic-based approach captured additional species outside of *E. coli*. We did not apply enrichment sequencing to soil and water, thus cannot compare findings related to the

environment. Together, these paired efforts highlight that multiple approaches may be needed to obtain a more complete understanding of AMR in a given context. Finally, our results are consistent with the paired study and UrbanZoo[11] in that when human to animal overlap is observed, it may occur between households.

We observed ARG sharing both through plasmids and through chromosomes. We note that *chromosome* was classified as sequences that were not identified as plasmids, and not through whole genome or strain analysis. Diverse mobile elements carrying ARGs have been observed in humans and animals. Horizontal gene transfer can facilitate transfer of AMR through these reservoirs. Previous work on *E. coli* in Nairobi concluded organismal spread, rather than transduction or transformation, was the dominant mechanism of highly similar mobile elements between human and animal hosts.[44] Similarly, the paired study of *E. coli* isolates in our same study households found strain sharing was more likely to contribute to resistome sharing than horizontal gene transfer.[16] Notably both studies were conducted on a single species, and our work demonstrates shared ARGs and genomic context between species (*E. coli* and *H. influenzae*, *K. pneumoniae* and *E. coli*) which likely occurred through horizontal gene transfer.

This work has several limitations. As previously mentioned, we processed samples on individual flow cells which is expensive and infeasible for a large study. Multiplexing samples has the potential to significantly reduce costs, though may result in decreased coverage. In addition, Oxford Nanopore Technologies' long-read sequencing is a relatively new technology and is consistently changing to improve the nominally high error rate (≈ 90-95% accuracy). This project was conducted on previous generation flow cells (R9.4.1), which are available from ONT upon request. Additional validation would need to be conducted for R10.4.1 since the duplex chemistry is a significant shift from the previous versions. Similar methods could be applied to alternative long-read technology such as PacBio.[47] Finally, we did not process environmental samples in this study; future work to process environmental samples (e.g. soil and water) with Context-Seq is recommended for a complete One Health approach to investigating AMR.[48]

Context-Seq is a promising approach for enriched long-read sequencing. While we demonstrate the utility of this assay with two ARGs, there is potential for ARG multiplexing, sample multiplexing, and further optimization of the enriched alleles. A previously published method (FLASH) used Cas9 with

short-read sequencing to target detection of 127 genes with 5513 guides.[14] Since many ARGs are co-located,[49] a guide pool of this size would likely be counterproductive to obtaining long reads but there is possible room for target expansion before compromising read length. One potential area of expansion is including guides for different alleles (e.g. CTX-M group 1 vs. CTX-M group 9) as they would likely be present on different DNA fragments. Further, the greatest potential for improving this method is sample multiplexing to reduce costs. While we ran each sample on a single flow cell, multiplexing on a MinION or promethION would significantly reduce the per sample cost. However, multiplexing is non-trivial and requires careful optimization to reduce off-target reads as it adds an additional step where non-target fragments can shear and introduce phosphorylated ends available for adapter ligation. Lower cost Context-Seq could be transformative to inform transmission dynamics of AMR through human, animal, and environmental reservoirs in diverse settings.

Methods

Sample Collection and Processing

Poultry-owning households from Dagoretti South and Kibera subcounties of Nairobi, Kenya were sampled in June-August 2019. Up to three poultry fecal samples and one canine fecal sample were collected during an initial visit. To collect animal feces, a sterile plastic scoop was used to transfer feces from the top, center layer of a fresh fecal pile. Approximately one week after the first visit, households were revisited to collect human stool from one household member in the following three age groups: child aged 0 - 4 years, child aged 5 - 14 years, and adult aged 15 years or older. A stool collection kit was provided during the first visit, which included a 50 mL plastic pot with a sterile scoop for each member with instructions on how to collect the sample. The primary caretaker of each household was informed by mobile phone one day prior to the revisit to collect stool from the previous night or the morning of the revisit day. All human and animal fecal samples were placed in a cooler filled with ice and transported to KEMRI. 1 g of fresh feces was aliquoted for storage at -80°C without preservatives. DNA was extracted from animal samples at Kenya Medical Research Institute (KEMRI) and stored at -80°C until transport. Human fecal samples and DNA extracts from animals were shipped to Tufts University on dry ice. For all

fecal samples, DNA was extracted from 0.2 g of feces using Qiagen's Powersoil Pro kit according to the manufacturer's instructions.

The study was approved by the KEMRI Scientific and Ethics Review Unit (Protocol number 3823) and the Tuft Health Sciences Institutional Review Board (13205). Additionally, a research permit was granted by the Kenyan National Commission for Science, Technology, and Innovation.

Taqman Array Card

Five adult fecal samples, 15 child fecal samples, four canine, and 15 poultry samples from eight households were prescreened for antibiotic resistance genes[50] using a Taqman Array Card prior to enrichment sequencing. 22 targets were run in duplicate, 14 of which were ARGs. Samples positive for CTX-M group 1 and the TEM assay were candidates for enrichment sequencing (**Table S2**) (see supporting information for more details).

gRNA Design

Candidate guide RNAs (gRNAs) for ARGs were determined by CHOPCHOP.[30] A custom program (https://github.com/Shruteek/Optimized-sgRNA-Design) was created to screen off-target effects in representative metagenomes. The program, implemented in Python, used empirical methods for on- and off-target effect analysis by taking in a candidate sequence and a sample metagenome, and returning a heuristic representing the overall likelihood of the candidate sequence to experience off-target effects in the metagenome. Based on guide RNA binding behavior,[31] we counted off-target sites as valid only if they had 5 or fewer mismatches, 1 or fewer mismatches in the 10 PAM-proximal base-pairs, and a PAM of the form 5'-NGG-3' or 5'-NRG-3'. Bowtie[51] was used to identify potential off-target sites, then each off-target site was evaluated for its binding likelihood based on the number of PAM sequences in the forward and reverse sequence,[52] the 1- and 2-base-pair nucleotide features in and around the site,[53] the identity of each nucleotide,[53–55] the individual nucleotide mismatches between the site and the guide,[31] and the proximity of each mismatch to the PAM.[31,56,57] The binding likelihood scores of the on-target sequence and each off-target site were normalized from 0 to 100, the latter corresponding to maximum

binding odds for a perfectly stable matching sequence, and all off-target likelihood scores for a single guide were summed and subtracted from the on-target likelihood score to generate the heuristic for off-target effects.

Library Preparation

We modified a previously published Cas9 enrichment protocol.[15] To evaluate performance and test modifications, we made a model system comprised of an *E. coli* isolate with $bla_{\text{CTX-M-55}}$ and $bla_{\text{TEM-1}}$ genes spiked into a composited DNA extracted from Kenyan soil (see supporting information for additional details). Unless otherwise specified, the protocol was performed as described below. For the evaluated modifications we made the following adjustments 1.) For adaptive sequencing, the MinKNOW software was set up in adaptive mode using the $bla_{\text{CTX-M-55}}$ gene as the reference and aligning up to 200 bps. Adaptive sequencing is a software-based method that allows the MinKNOW software to read the first few hundred base pairs of a fragment and selectively reject the fragment from the pore if it is classified as off-target.[58] 2.) For longer Cas9 cut time, Cas9 digestion proceeded for 2 hours instead of 20 minutes. 3.) For two guides per target per strand (sense and antisense), guides were added in an equimolar mix of 0.75 µL each to a 0.5 mL centrifuge tube and 1uL of the mix was complexed with Cas9. The protocol below describes the addition of Proteinase K as that was incorporated in our final procedure.

CrRNAs and tracrRNAs, together forming the gRNA, were resuspended to a final concentration of 100 µM in duplex buffer (IDT). 8 µL of nuclease free water ,1 µL of tracR, and 1 µL of crRNA were mixed and heated at 95°C for 5 minutes for duplex formation. To create the ribonucleoprotein complex (RNP), 1X CutSmart Buffer (NEB), 2 µM of gRNA, 0.5 µM of HiFi Cas9 Nuclease V3 (IDT), and nuclease free water were combined to a total reaction volume of 30 µL. The reaction was incubated at room temperature for 20 minutes. Input DNA (approximately 1.5-3.0 µg) was dephosphorylated in a 60 µL reaction composed of 6 µL 1X CutSmart buffer, DNA, nuclease free water, and 3 µL of QuickCIP (NEB). The reaction was incubated at 27°C for 20 minutes followed by inactivation at 80°C for 2 minutes. For Cas9 cleavage and A-tailing, dephosphorylated DNA was split up into two reactions (2 reactions of 30 µL). Input DNA was split to allow for Cas9 cutting on the sense and antisense strands separately using two sets of guide RNAs for the two target ARGs. 30 µL of DNA and 10 µL of RNP were mixed and

incubated at 37°C for 15 minutes. 1µL thermolabile Proteinase K (NEB) was added and incubate at 37 °C for 10 min. Proteinase K was then heat inactivated at 65 °C for 10 minutes. 1 µL dATP (Invitrogen) and 1µL Taq polymerase (NEB) were added to the mixture and incubated at 37 °C for 15 min followed by 72 °C for 5 minutes. To ligate on sequencing adapters, ligation mix was prepared by adding 9 µL nuclease free water, 40 µL ligation buffer, 20 µL T4 quick ligase (NEB), and 7 µL of adapters. 38 µL of the adapter mix was added to each reaction and incubated at room temperature for 10 minutes on a tube rotator. Equal volume of TE buffer was added to each reaction and then the two libraries (one for sense and another for antisense strand) were pooled. 0.5X Ampure beads (Beckman Coulter) were added. The reaction was rotated for 5 minutes followed by incubation at room temperature on a bench top for 5 minutes. Magnetic beads were washed with 250 µL of long fragment buffer. After addition of 13 µL of elution buffer, beads were incubated at 37°C for 30 minutes. MinION flow cells were loaded according to the manufacturer's instructions with 12 µL DNA in elution buffer, 25.5 µL loading beads, and 37.5 µL sequencing buffer.

Sequencing

Samples were run on MinION (FLO-MIN106) R9.4.1 flow cells using a MK1B sequencer. Runs were operated using MinKNOW software (v22.05.5, v22.10.10, v22.12.7, v23.04.6).

Data Analysis

Fast5 files were basecalled using guppy (v6.1.5, v6.3.9, v6.4.6, v6.5.7) with a minimum quality score of 7. Porechop (v0.2.4)[59] was used to trim remaining nanopore sequencing adapters. Usearch (v11.0.667)[60] was used to sort trimmed reads by length and cluster reads with a minimum overlap of 1500 bp at 85% identity. Three cycles of racon (v1.4.20)[51] followed by medaka (0.11.5)[61] were used to polish centroids with the reads assigned to the same cluster, generating one consensus sequence per cluster. Singletons were included without polishing. Antibiotic resistance genes (ARGs) were identified in resulting consensus sequences (including singletons) using Minimap2 (2.22-r1101)[62] to map against the Comprehensive Antibiotic Resistance Database (CARD) (v3.2.6).[63] Mobile genetic elements (MGEs) were

annotated using the mobileOG database (beatrix-1.6). The mobileOG[64] database is a manual curation of MGEs from ICEBerg, ACLAME, GutPhage Database, Prokaryotic viral orthologous groups, COMPASS, NCBI Plasmid RefSeq, immedb, and ISfinder, along with homologs of the manually curated sequences. Consensus sequences were also annotated using Prokka (v1.14.5).[65] Taxonomy was assigned to consensus sequences using Kraken2 (v2.0.7-beta)[66] with the default full database. BLASTN webserver was used to map consensus sequences against NCBI core non-redundant nucleic acid database (core_nt). PlasX[67] was used to identify the probability contigs were plasmid sequences. Sequences with a PlasX score > 0.5 (0 likely not plasmid, 1 likely plasmid) were labelled as plasmids. BLASTn (2.12.0) all-versus-all was used to identify regions of overlap between samples at greater than 80% identity for 3000 bps. Annotated consensus sequences were visualized in R (4.3.2) using ggenes (0.5.1). Coverage plots for benchmarking were visualized using genomicRanges (1.54.1) and genomicAlignments (1.38.0).

Acknowledgements

This work was supported by grant OPP1129535 from the Bill and Melinda Gates Foundation and the Chan Zuckerberg Biohub, San Francisco. Support also came from the National Center for Advancing Translational Sciences, National Institutes of Health, Award Number UL1TR002544 and the Stuart B. Levy Center for Integrated Management of Antimicrobial Resistance at Tufts (Levy CIMAR), a collaboration of Tufts Medical Center and the Tufts University Office of the Vice Provost for Research (OVPR) Research and Scholarship Strategic Plan (RSSP). ERF was supported by the NSF Postdoctoral Research Fellowships in Biology Program under Grant No. 1906957. We thank Maya Nadimpalli, Robert Gilman, and Monica Pajeulo for their contribution of the positive control isolate (NIH R01AI108695-01A1 and Tufts Springboard award). We thank Honey Mekonen, Ritwicq Arjyal, Joana Cabrera, and Andres Dextre for research assistance. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding organizations.

Author contributions

Project conceptualization was performed by ERF and AJP. Methodology for this work was developed by ERF, SK, SAM, AP. Samples were collected by JMS and BC. Laboratory experiments were performed by ERF, SK, CM, and AP. Data analysis was conducted by ERF, SAM, SK, and CM. Visualization of data and results was performed by ERF. Funding for this work was acquired by ERF, AJP, and SMN. Writing of original draft was carried out by ERF, SAM, and AJP. Reviewing and editing of the manuscript was performed by all.

Competing interests

The authors declare no competing interests.

Data Availability

All fastq files are available in the Sequence Read Archives under BioProject PRJNA1157857.

Code Availability

All codes for the guide design tool are available at https://github.com/Shruteek/Optimized-sgRNA-Design.

References

1. Murray, C. J. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629–655 (2022).
2. Fuhrmeister, E. R. *et al.* Evaluating the relationship between community water and sanitation access and the global burden of antibiotic resistance: an ecological study. *The Lancet Microbe* **4**, e591–e600 (2023).
3. Nadimpalli, M. L. *et al.* Urban informal settlements as hotspots of antimicrobial resistance and the need to curb environmental transmission. *Nat Microbiol* **5**, 787–795 (2020).
4. Klein, E. Y. *et al.* Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. *PNAS* **115**, E3463–E3470 (2018).

5. Huijbers, P. M. C. *et al.* Role of the Environment in the Transmission of Antimicrobial Resistance to Humans: A Review. *Environ. Sci. Technol.* **49**, 11993–12004 (2015).

6. Delgado-Baquerizo, M. *et al.* The global distribution and environmental drivers of the soil antibiotic resistome. *Microbiome* **10**, 219 (2022).

7. Li, J. *et al.* Global Survey of Antibiotic Resistance Genes in Air. *Environ. Sci. Technol.* **52**, 10975–10984 (2018).

8. Zhu, Y.-G. *et al.* Continental-scale pollution of estuaries with antibiotic resistance genes. *Nature Microbiology* **2**, 16270–16270 (2017).

9. Brito, I. L. Examining horizontal gene transfer in microbial communities. *Nat Rev Microbiol* **19**, 442–453 (2021).

10. Montealegre, M. C. *et al.* High Genomic Diversity and Heterogenous Origins of Pathogenic and Antibiotic-Resistant Escherichia coli in Household Settings Represent a Challenge to Reducing Transmission in Low-Income Settings. *mSphere* **5**, e00704-19.

11. Muloi, D. M. *et al.* Population genomics of Escherichia coli in livestock-keeping households across a rapidly developing urban landscape. *Nat Microbiol* **7**, 581–589 (2022).

12. Salinas, L. *et al.* Environmental Spread of Extended Spectrum Beta-Lactamase (ESBL) Producing Escherichia coli and ESBL Genes among Children and Domestic Animals in Ecuador. *Environmental Health Perspectives* **129**, 027007.

13. Vikesland, P. J. *et al.* Toward a Comprehensive Strategy to Mitigate Dissemination of Environmental Sources of Antibiotic Resistance. *Environmental Science & Technology* **51**, 13061–13069 (2017).

14. Quan, J. *et al.* FLASH: a next-generation CRISPR diagnostic for multiplexed detection of antimicrobial resistance sequences. *Nucleic acids research* **47**, e83 (2019).

15. Gilpatrick, T. *et al.* Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature Biotechnology* **38**, 433–438 (2020).

16. Kim, D. D. *et al.* Bacterial strain sharing between humans, animals, and the environment among urban households. 2024.08.05.24311509 Preprint at https://doi.org/10.1101/2024.08.05.24311509 (2024).

17. Zhang, A.-N. *et al.* An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nat Commun* **12**, 4765 (2021).

18. Bradford, P. A. Extended-Spectrum β-Lactamases in the 21st Century: Characterization, Epidemiology, and Detection of This Important Resistance Threat. *Clin Microbiol Rev* **14**, 933–951 (2001).

19. Bevan, E. R., Jones, A. M. & Hawkey, P. M. Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype. *Journal of Antimicrobial Chemotherapy* **72**, 2145–2155 (2017).

20. Woerther, P.-L., Burdet, C., Chachaty, E. & Andremont, A. Trends in Human Fecal Carriage of Extended-Spectrum β-Lactamases in the Community: Toward the Globalization of CTX-M. *Clinical Microbiology Reviews* **26**, 744–758 (2013).

21. Langata, L. M., Maingi, J. M., Musonye, H. A., Kiiru, J. & Nyamache, A. K. Antimicrobial resistance genes in Salmonella and Escherichia coli isolates from chicken droppings in Nairobi, Kenya. *BMC Res Notes* **12**, 22 (2019).

22. Borgogna, T. R. *et al.* High Diversity of CTX-M Extended-Spectrum β-Lactamases in Municipal Wastewater and Urban Wetlands. *Microb Drug Resist* **22**, 312–320 (2016).

23. Gatica, J. *et al.* Resistance of Undisturbed Soil Microbiomes to Ceftriaxone Indicates Extended Spectrum β-Lactamase Activity. *Front. Microbiol.* **6**, (2015).

24. Cantón, R., González-Alba, J. M. & Galán, J. C. CTX-M Enzymes: Origin and Diffusion. *Front Microbiol* **3**, 110 (2012).

25. Cloeckaert, A. *et al.* Dissemination of an Extended-Spectrum-β-Lactamase blaTEM-52 Gene-Carrying IncI1 Plasmid in Various Salmonella enterica Serovars Isolated from Poultry and Humans in Belgium and France between 2001 and 2005. *Antimicrobial Agents and Chemotherapy* **51**, 1872–1875 (2007).

26. Singh, N. S., Singhal, N. & Virdi, J. S. Genetic Environment of blaTEM-1, blaCTX-M-15, blaCMY-42 and Characterization of Integrons of Escherichia coli Isolated From an Indian Urban Aquatic Environment. *Front. Microbiol.* **9**, (2018).

27. Gundran, R. S. *et al.* Prevalence and distribution of blaCTX-M, blaSHV, blaTEM genes in extended- spectrum β- lactamase- producing E. coli isolates from broiler farms in the Philippines. *BMC Veterinary Research* **15**, 227 (2019).

28. Lachmayr, K. L., Kerkhof, L. J., Dirienzo, A. G., Cavanaugh, C. M. & Ford, T. E. Quantifying nonspecific TEM beta-lactamase (blaTEM) genes in a wastewater stream. *Appl Environ Microbiol* **75**, 203–211 (2009).

29. Mabilat, C. & Courvalin, P. Development of 'oligotyping' for characterization and molecular epidemiology of TEM beta-lactamases in members of the family Enterobacteriaceae. *Antimicrobial Agents and Chemotherapy* **34**, 2210–2216 (1990).

30. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Research* **47**, W171–W174 (2019).

31. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**, 827–832 (2013).

32. Anjum, M. F. *et al.* The potential of using *E. coli* as an indicator for the surveillance of antimicrobial resistance (AMR) in the environment. *Current Opinion in Microbiology* **64**, 152–158 (2021).

33. Ashurst, J. V. & Dawson, A. Klebsiella Pneumonia. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2024).

34. Effah, C. Y., Sun, T., Liu, S. & Wu, Y. Klebsiella pneumoniae: an increasing threat to public health. *Ann Clin Microbiol Antimicrob* **19**, 1 (2020).

35. Wyres, K. L., Lam, M. M. C. & Holt, K. E. Population genomics of Klebsiella pneumoniae. *Nat Rev Microbiol* **18**, 344–359 (2020).

36. WHO bacterial priority pathogens list, 2024: Bacterial pathogens of public health importance to guide research, development and strategies to prevent and control antimicrobial resistance. https://www.who.int/publications/i/item/9789240093461.

37. Muraya, A. *et al.* Antimicrobial Resistance and Virulence Characteristics of Klebsiella pneumoniae Isolates in Kenya by Whole-Genome Sequencing. *Pathogens* **11**, 545 (2022).

38. Chepkemei, A., Mwaniki, J., Nyerere, A. & Kiiru, J. Phenotypic and Genotypic Characterisation of Antibiotic Resistance in Escherichia coli, Klebsiella spp., and Listeria monocytogenes Isolates from Raw Meat Sold in Nairobi. *Advances in Microbiology* **12**, 603–620 (2022).

583 39. Slack, M. P. E. A review of the role of Haemophilus influenzae in community-acquired
584 pneumonia. *Pneumonia* **6**, 26–43 (2015).
585 40. Bae, S. *et al.* Antimicrobial Resistance in Haemophilus influenzae Respiratory Tract Isolates
586 in Korea: Results of a Nationwide Acute Respiratory Infections Surveillance. *Antimicrob*
587 *Agents Chemother* **54**, 65–71 (2010).
588 41. Torumkuney, D. *et al.* Results from the Survey of Antibiotic Resistance (SOAR) 2015–18 in
589 Tunisia, Kenya and Morocco: data based on CLSI, EUCAST (dose-specific) and
590 pharmacokinetic/pharmacodynamic (PK/PD) breakpoints. *Journal of Antimicrobial*
591 *Chemotherapy* **75**, i2–i18 (2020).
592 42. Muloi, D. *et al.* A cross-sectional survey of practices and knowledge among antibiotic
593 retailers in Nairobi, Kenya. *J Glob Health* **9**, 010412 (2019).
594 43. Kemp, S. A., Pinchbeck, G. L., Fèvre, E. M. & Williams, N. J. A Cross-Sectional Survey of the
595 Knowledge, Attitudes, and Practices of Antimicrobial Users and Providers in an Area of
596 High-Density Livestock-Human Population in Western Kenya. *Front Vet Sci* **8**, 727365 (2021).
597 44. Hassell, J. M. *et al.* Clinically relevant antimicrobial resistance at the wildlife–livestock–
598 human interface in Nairobi: an epidemiological study. *The Lancet Planetary Health* **3**, e259–
599 e269 (2019).
600 45. Swarthout, J. M. *et al.* Differential Overlap in Human and Animal Fecal Microbiomes and
601 Resistomes in Rural versus Urban Bangladesh. *Applied and Environmental Microbiology* **88**,
602 e00759-22 (2022).
603 46. Swarthout, J. M., Chan, E. M. G., Garcia, D., Nadimpalli, M. L. & Pickering, A. J. Human
604 Colonization with Antibiotic-Resistant Bacteria from Nonoccupational Exposure to

Domesticated Animals in Low- and Middle-Income Countries: A Critical Review. *Environ. Sci. Technol.* **56**, 14875–14890 (2022).

47. Wieben, E. D. *et al.* Amplification-free long-read sequencing of TCF4 expanded trinucleotide repeats in Fuchs Endothelial Corneal Dystrophy. *PLoS One* **14**, e0219446 (2019).

48. Davis, M. F. *et al.* Checklist for One Health Epidemiological Reporting of Evidence (COHERE). *One Health* **4**, 14–21 (2017).

49. Li, B. *et al.* Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *The ISME Journal* **9**, 2490–2502 (2015).

50. Pholwat, S. *et al.* Genotypic antimicrobial resistance assays for use on E. coli isolates and stool specimens. *PLOS ONE* **14**, e0216747 (2019).

51. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).

52. Malina, A. *et al.* PAM multiplicity marks genomic target sites as inhibitory to CRISPR-Cas9 editing. *Nat Commun* **6**, 10124 (2015).

53. Moreno-Mateos, M. A. *et al.* CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Methods* **12**, 982–988 (2015).

54. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).

55. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**, 1262–1267 (2014).

56. Anderson, E. M. *et al.* Systematic analysis of CRISPR–Cas9 mismatch tolerance reveals low levels of off-target activity. *Journal of Biotechnology* **211**, 56–65 (2015).

57. Mendoza, B. J. & Trinh, C. T. Enhanced guide-RNA design and targeting analysis for precise CRISPR genome editing of single and consortia of industrially relevant and non-model organisms. *Bioinformatics* **34**, 16–23 (2018).
58. Martin, S. *et al.* Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome Biology* **23**, 11 (2022).
59. Wick, R. rrwick/Porechop. (2024).
60. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
61. Medaka. Oxford Nanopore Technologies (2022).
62. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
63. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* **48**, D517–D525 (2020).
64. Brown, C. L. *et al.* mobileOG-db: a Manually Curated Database of Protein Families Mediating the Life Cycle of Bacterial Mobile Genetic Elements. *Applied and Environmental Microbiology* **88**, e00991-22 (2022).
65. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
66. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257 (2019).

647 67. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their ecology across
648 human gut metagenomes revealed by PlasX and MobMess. *Nat Microbiol* **9**, 830–847
649 (2024).
650