

1 **Diversity of Antibiotic Resistance genes and Transfer Elements-Quantitative Monitoring**

2 **(DARTE-QM): a method for detection of antimicrobial resistance in environmental samples**

3

4 **Schuyler D. Smith**

5 *Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA*

6 *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA*

7 Email: sdsmit@iastate.edu

8

9 **Jinlyung Choi**

10 *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA*

11 Email: genase23@gmail.com

12

13 **Nicole Ricker**

14 Previous: *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA;*

15 *Food Safety and Enteric Pathogens Research Unit, ARS-USDA National Animal Disease Center, Ames,*

16 *IA*

17 Current: *Department of Pathobiology, Ontario Veterinary College, University of Guelph, Guelph, ON,*

18 *Canada*

19 Email: nricker@uoguelph.ca

20

21 **Fan Yang**

22 *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA*

23 Email: fan.michelle.yang@gmail.com

24

25 **Shannon Hinsa-Leasure**

26 *Department of Biology, Grinnell College, Grinnell, IA*

27 Email: hinsa@grinnell.edu

28

29 **Michelle Soupir**

30 *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA*

31 Email: msoupir@iastate.edu

32

33 **Heather Allen**

34 *Food Safety and Enteric Pathogens Research Unit, ARS-USDA National Animal Disease Center, Ames,*

35 *IA*

36

37 **Adina Howe***

38 *Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA*

39 *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA*

40 Email: adina@iastate.edu

41

42 ABSTRACT

43

44 Effective monitoring of antibiotic resistance genes and their dissemination in environmental
45 ecosystems has been hindered by the cost and efficiency of methods available for the task. We
46 developed a method entitled the Diversity of Antibiotic Resistance genes and Transfer Elements-
47 Quantitative Monitoring (DARTE-QM), a system implementing high-throughput sequencing to
48 simultaneously sequence thousands of antibiotic resistant genes representing a full-spectrum of
49 antibiotic resistance classes commonly seen in environmental systems. In this study, we demonstrated
50 DARTE-QM by screening 662 antibiotic resistance genes within environmental samples originated
51 from manure, soil, and animal feces, in addition to a mock-community used as a control to test
52 performance. DARTE-QM offers a new approach to studying antibiotic resistance in environmental
53 microbiomes, showing advantages in efficiency and the ability to scale for many samples. This method
54 provides a means of data acquisition that will alleviate the obstacles that many researchers in this area
55 currently face.

56

57

58 INTRODUCTION

59

60 The global spread of organisms possessing antimicrobial resistance (AMR), and their associated
61 antibiotic resistant genes (ARGs), is posing an increasing threat to the health of both humans and
62 animals alike¹⁻³. Characterization of the presence and abundance of ARGs, i.e. the resistome, in
63 environmental microbiome samples has stood as a major challenge for researchers monitoring these
64 events⁴. Such studies have been impeded by the broad diversity of the genes, their low presence in most
65 natural environments, the difficulty of extracting DNA from microbes in those environments, and their

66 association with mobile genetic elements accounting for approximately one-quarter of the genetic
67 material in these microbiomes⁵.

68

69 The genetic diversity of ARGs has made targeted sequencing approaches non-trivial and has led to the
70 application of whole-genome shotgun metagenomic methods for the characterization resistomes⁶. This
71 approach is dependent on the availability of a gene reference database to classify reads as ARGs
72 sequences but does not require *a priori* knowledge of which genes constitute the resistome being
73 investigated⁷. Despite being effective for the task, the cost per sample of employing metagenomic
74 methods to elucidate resistomes often inhibits studies from scaling. Shotgun sequencing must
75 indiscriminately sequence a genome, and often the resistome comprises only a fraction of a percent of
76 the entire metagenome. Therefore, it is often the case that only a minute subset of the sequencing-reads
77 produced through this method will be informative to resistomes, and ARGs are either underrepresented
78 or undetected⁸, as sufficient sequencing depth and coverage is difficult to achieve.

79

80 In the effort to find more efficient means for sequencing ARGs, a method of implementing bait-and-
81 capture system to identify ARG targets has been developed⁹. This approach uses streptavidin-coated
82 magnetic beads to capture 80-mer bait sequences to target genes of interest. The bait-and-capture
83 method has been well-suited for the characterization of low- and high-abundance ARGs and has
84 demonstrated the ability to differentiate resistomes from different sample sources¹⁰. Another method of
85 targeted gene sequencing used for ARG characterization involves custom primers for performing a
86 PCR-based amplicon library preparation. This type of sequencing is used extensively in microbiome
87 studies for community profiling via bacterial 16S rRNA genes and combines barcoded adapters to
88 differentiate hundreds of samples pooled in a single library preparation¹¹. It has previously been limited
89 in the number of primers that could be incorporated for a single library, but a more recent version of

90 amplicon library preparation for multiplexed primers now exists and has been implemented for
91 biomarker detection in clinical studies¹²⁻¹⁵.

92

93 Our study demonstrates the first usage of this multiplexed amplicon library preparation for the
94 detection of ARGs in environmental samples. We have termed our method of implementing this
95 technology Diversity of Antibiotic Resistance genes and Transfer Elements-Quantitative Monitoring
96 (DARTE-QM). Our study was designed to demonstrate that DARTE-QM offers practical application to
97 ARG screening through its ability to simultaneously detect and quantify hundreds of ARGs residing in
98 samples from various environments and that it can achieve high accuracy and sensitivity identifying
99 ARG targets.

100

101 RESULTS

102

103 **Design of primers and samples.** DARTE-QM employed 796 primer pairs designed to target 67
104 antibiotic resistant families and 662 ARGs, as well as a synthetic oligonucleotide reference sequence
105 and the V4 region of the 16S rRNA gene, in a multiplexed amplicon library preparation (Supp. Table
106 1, Supp. Table 2). Subsequent paired-end sequencing of 150 base pair reads was conducted using the
107 Illumina MiSeq platform (USDA, Ames, IA). To evaluate the results of DARTE-QM against a
108 reference, we constructed a mock-community microbiome comprised of DNA extracted from 20
109 isolates (Supp. Table 3) with completed genome sequences (Dataset 1). For each of the mock-
110 community libraries, we included varying concentrations of a synthetic oligonucleotide reference
111 sequence to evaluate accuracy of quantification. We also examined how DARTE-QM was able to
112 characterize true environmental resistomes associated with manure, swine fecal, and agricultural soil
113 samples (Supp. Table 4).

114

115 **Evaluation of DARTE-QM sequencing products.** The sequencing data produced via DARTE-QM is
116 unique in its high level of heterogeneity, as compared to traditional amplicon data generated from a
117 singular DNA-primer (e.g., 16S SSU rRNA). Given numerous and diverse gene targets in the
118 sequencing library, processing of DARTE-QM data required amendment of the traditional microbiome
119 analysis pipelines (Figure 1). After quality control and processing, 16 of the 18 samples from the mock-
120 community were retained for downstream analysis (2 samples removed for less than 5,000 reads
121 passing quality filters). Quality filtering also resulted in the removal of 38 of the 61 environmental
122 samples due to sequencing coverage below 5000 reads, likely caused by PCR inhibitors common of
123 manure and soil samples^{16–18}, leaving 39 samples in total to be used in the evaluation of DARTE-QM.
124 The 16 mock-community samples yielded a mean of 192,415 reads per sample, and a mean of 44,440
125 reads able to be aligned to ARG references (Supp. Table 5). In our environmental samples, across all
126 sources, we observed a mean of 170,775 reads and a mean of 19,138 reads aligned to ARG references
127 per sample.

128

129 **DARTE-QM successfully amplified targeted genes with high accuracy and sensitivity.** Reads from
130 each sample were demultiplexed by primer, and each read was subsequently classified as either true
131 positive (TP), false positive (FP), false negative (FN), or true negative (TN). This classification was
132 based on alignment to the ARG reference database, where reads were deemed to be TP when both the
133 intended primer target and read sequence aligned to the same gene; FP when the primer target and the
134 read sequence did not agree; FN when no primer was found but the read sequence was able to be
135 aligned to a reference ARG; and TN reads were assigned as all reads within a sample assigned as TP
136 outside of the primer in question, i.e., all reads that were correctly identified as not being the targeted
137 read.

138 Success for DARTE-QM was evaluated on three metrics: sensitivity (TP/[TP + FN]), specificity
139 (TN/[TN + FP]), and accuracy ([TP + TN]/[TP + FN + TN + FP]) for each gene target (i.e., primer) and
140 each sample (Supp. Table 6). From the 662 ARGs targeted by DARTE-QM, 235 (~35%) were identi-
141 fied in our samples. The mean sensitivity for all primers was found to be 99.6%. The mean specificity
142 and mean accuracy were found to be > 99.9% and 99.6%, respectively, suggesting that the primers in
143 DARTE-QM were successful in amplifying their intended target genes.

144

145 We also observed a substantial number of reads in our sequencing libraries that had primers located on
146 the 5' end of the sequences but were unable to be aligned to any of our reference ARGs nor any posi-
147 tion in the mock-community genomes. Inspection of a subset of these reads found that they contained
148 repeated poly-A and poly-T elements. These reads were observed as unique sequences within the da-
149 taset, implying little or no biological pattern. These artifacts accounted for 47% of all reads in samples
150 which passed quality controls. However, in samples that failed to pass quality filters, these sequences
151 accounted for 85% of reads. Sample source appeared to be a significant, yet likely confounded, factor
152 in the production of these artifacts. Sequencing of samples from the mock-community had significantly
153 lower counts for artifact reads as compared to environmental samples (soil-A, $p = 0.038$; manure-A +
154 soil-A, $p = 0.035$; swine fecal, $p < 0.001$, pairwise-Wilcoxon). Across all samples, an inverse linear re-
155 lationship ($R^2 = 0.68$) (Supp. Figure 1) was observed between the number of reads which had a primer
156 identified and the percentage of those reads that were artifacts.

157

158 **DARTE-QM was able to consistently identify presence and distribution of ARGs.** Construction of
159 the mock-community from DNA sourced from fully-sequenced genomes allowed for comparison of a
160 theoretical profile to our experimental observations of ARGs in these samples. In the combined
161 genomes of the mock-community, ARGs comprised 0.03% (56 ARG targets) of the total genome by

162 base pair count. DARTE-QM was able to produce 55 of those 56 ARGs found in the mock-community
163 reference genomes, consistently identifying them across all 16 samples (Figure 2). Particular resistance
164 families that were not successfully captured by DARTE-QM included those associated with the *acrA*
165 subunit of multidrug efflux pump systems, as well as genes encoding for chloramphenicol resistance
166 (e.g., *catA*). While overall, target relative abundances were observed to be similar compared to
167 theoretical, the quantification of particular ARGs, such as transposon-associated *lnuC* conferring
168 resistance to lincomycin, were found in higher abundance by DARTE-QM, as others such as *mecA*
169 conferring methicillin-resistance, were found to be underrepresented. With regard to the synthetic
170 oligonucleotide, there was a strong correlation observed reads (Supp Figure 2, $R^2 = 0.91$) between the
171 read abundance produced by DARTE-QM and the experimental concentration.

172

173 **DARTE-QM differentiated resistomes between environmental sources.** DARTE-QM detected 240
174 ARG targets across all samples in this study (including 121 in Soil-A, 172 in Soil-B, 182 in Soil-C, 202
175 in Swine Manure-A, 129 in Swine Manure-B, 178 in the Swine Fecal samples, and 156 in the mock-
176 community, Supp. Table 4). Distinctions in the composition of resistomes were detected, not just from
177 the presence of unique ARG targets but also from the abundance of the ARGs that composed the
178 resistomes from each environment (Figure 3a). Ordination, via principal coordinate analysis based on
179 Bray-Curtis distances of observed ARGs targets, showed clear separation of environmental sources,
180 with the first two eigenvalues accounting for nearly 80% of the total variation (Figure 3b).
181 Permutational multivariate analysis of variance (PERMANOVA) was used as a non-parametric
182 multivariate statistical test to compare the variation of samples and environmental source. The results
183 of the PERMANOVA test corroborated the apparent findings of the PCoA, and environmental sources
184 were associated with a significant ($F=11.45$, $R^2=0.70$, $p < 0.001$) portion of variation observed in the
185 resistome profiles. DARTE-QM identified specific ARG patterns which distinguished resistomes

186 sourced from different environmental samples, the most notable of which was within swine fecal
187 samples where a distinctive presence of genes related to lincosamide and aminoglycoside resistance
188 were observed. In the soils, with varied field management histories of swine and bovine manure
189 amendment (soils B and C), we observed distinct characteristics of resistomes as well. Bovine manure-
190 associated soils were found to be enriched with genes associated with resistance to aminoglycosides
191 and sulfanomides, whereas the swine manure-amended soils were replete with aminoglycoside,
192 lincosamides, and erythromycin-resistance related genes.

193

194 **DARTE-QM produced results with comparable resolution to that of metagenomes.** Soil-column
195 samples used in this study had been previously characterized through metagenome sequencing¹⁹ (NCBI
196 SRA Study SRP193066). DNA from the same sources were used for sequencing with DARTE-QM
197 study for comparison of the two methods. Metagenomes from the soil samples had an average of 241
198 ARG reads and were excluded from analysis; DARTE-QM returned a mean abundance of 5,839 ARG
199 reads in those same samples. Four swine-manure samples from the metagenome study yielded a mean
200 ARG abundance of 76,226 reads, and the 12 manure treated soil samples yielded an average of 7,377
201 ARG reads. DARTE-QM produced mean abundances of 32,678 and 13,488 ARG reads in the same
202 samples.

203

204 Relative abundance of ARG classes showed similar profiles for swine-manure from both technologies.
205 DARTE-QM reads were classified into 99 ARG families and metagenome reads to 56 ARG families.
206 From those, 39 ARG families were shared between the two methods and accounted for 89% and 84%
207 of metagenome and DARTE-QM ARG families, respectively (Figure 4). In the manure-treated soil
208 samples DARTE-QM identified 99 ARG families and metagenomes 92, sharing 50 of those that ac-
209 counted for 90% and 83%, respectively. For identifying diverse ARGs, DARTE-QM is disadvantaged

210 by being a targeted method. For example, the metagenomes had a noticeable presence of genes from
211 the AMR gene families for resistance-nodulation-cell division antibiotic efflux pump (*Mux* and *Mex*
212 ARG Classes), which were not targeted by DARTE-QM. A direct comparison of both approaches con-
213 strains the metagenomes to those targeted by the primers of DARTE-QM (Figure 4b). In this compari-
214 son, metagenomes identified 48 ARG families in the swine manure samples and 65 in the manure-
215 treated soils. Diversity measurements using the Shannon-Weiner Index of ARG classes showed similar
216 values between the methods with DARTE-QM having $H = 2.95$ in swine-manure samples and $H = 2.87$
217 in manure-treated soil samples, while metagenomes had $H = 2.92$ in swine-manure samples and $H =$
218 2.84 in the manure-treated soils.

219

220 **DARTE-QM can distinguish gene variants through sequencing.** Two high-abundance genes, *erm35*
221 encoding for the macrolide-lincosamide-streptogramin and *tetM* for tetracycline resistance, were
222 selected for variant analysis. DARTE-QM reads classified as either of these genes were clustered at
223 97% nucleotide identity, resulting in three clusters for *erm35* and five clusters for *tetM*. Each cluster
224 contained a minimum of ten unique sequences. The primary *erm35* cluster contained 4,785 reads
225 (Supp. Table 6, Supp. Figure 3a). The other two *erm35* clusters were defined by 5 to 10 base pair
226 variations within the associated 13 and 18 reads. Similarly, from a total of 24,653 reads classified as
227 *tetM*, 96% defined the primary cluster, which was identical to one of the 6 *tetM* primer targets. Four of
228 the other clusters, which contained between 32 and 676 reads, were defined by 9 and 24 base pair
229 variations (Supp. Table 6, Supp Figure 3b). Bacterial hosts associated with the observed *erm35* variants
230 were solely associated with *Bacteroides coprosuis* and *Bacteroides spp.* and is consistent with the
231 limited diversity of known isolates carrying this gene. In contrast, the sequences associated with *tetM*
232 clusters are known to originate in various taxa. The largest *tetM* cluster was found to be highly
233 conserved across a broad diversity of Gram-positive and some Gram-negative isolates. In comparison,

234 the *tetM* cluster containing 676 reads, was primarily associated with plasmids found in *E. coli* and
235 *Salmonella*. The lower abundance of this cluster in the DARTE-QM data is consistent with the low
236 relative abundance of Enterobacterales in swine gut-associated samples²⁰. Similarly, the other *tetM*
237 clusters were associated with *Streptococcus* strains and found in a lower diversity of taxa compared
238 with the largest cluster.

239

240 DISCUSSION

241

242 DARTE-QM was conceptualized as an approach towards more efficient characterization of ARGs
243 found in microbiomes. Specifically, we developed DARTE-QM to address the cost limitations of
244 metagenomic approaches for ARG monitoring in environmental samples, where ARGs of interest often
245 require significant sequencing depth and coverage. One of the major goals was to drastically scale the
246 number of samples able to be evaluated by leveraging the high-throughput capabilities of barcode-
247 multiplexing combined with amplicon library preparation. Similar to other amplicon-sequencing
248 platforms, the costs of DARTE-QM are driven by the synthesis of primers and the price of sequencing.
249 As DARTE-QM targets specified genes for amplification, it is able to enrich and detect ARGs that are
250 present in low abundance, which is often a barrier for shotgun metagenomics. The number of samples
251 that can be processed using DARTE-QM is limited by the number of unique barcode sequence
252 adapters, the sequencing depth required per sample, and the number of gene targets. At the time of this
253 study, the number of gene targets was constrained by the TruSeq platform, which currently supports
254 1,536 primers and 96 barcoded samples.

255

256 The aim of this study was to demonstrate the efficacy of DARTE-QM for characterizing ARGs from
257 environmental samples. Our results showed that DARTE-QM had success detecting the presence of

258 hundreds of diverse ARGs across soil, manure, water, and our mock-community samples. While
259 DARTE-QM was designed with the capacity to identify diverse ARG targets, our assessment was
260 limited by ARGs contained in our samples. We used DNA extracted from isolates with known genomes
261 and ARG distributions to evaluate the sensitivity and accuracy of DARTE-QM. We observed strong
262 performance for detecting ARGs in our mock-community, having 98% of ARGs detected with high
263 sensitivity and specificity. There was evidence of DARTE-QM's ability to quantify ARG presence with
264 the correlation of abundance to varying concentrations of our synthetic oligonucleotide reference in the
265 mock-community samples. Those results, though not a perfect correlation ($r^2 = 0.91$), illustrate that
266 DARTE-QM is affected more by the amount of DNA available for the primer than by the competition
267 between primers to find targets. Finally, comparisons to metagenomes suggested that DARTE-QM
268 could detect similar measures of diversity of ARGs from samples. While the distributions of ARGs
269 within the resistomes varied between DARTE-QM and metagenome resistomes, the differences
270 between environmental sources could be distinguished, and broad patterns of resistance classes were
271 similar. Combined, these results confirm that the primers used for DARTE-QM successfully amplified
272 ARGs despite the potential for interference when simultaneously amplifying multiple gene targets in
273 uniform conditions.

274

275 In cases where DARTE-QM abundances varied the most from expected, the gene targets were often
276 associated with plasmids and other mobile elements. Multiple copies of these genes may exist per cell
277 and result in the underestimation of these genes. For instance, *aph3-ib*, *aph6-id* and *sul2* are found on
278 the same IncQ plasmid. This is a likely reason for the results of much higher observed copy numbers
279 than other ARGs, as well as the theoretical estimate. The IncQ plasmid has been reported to have
280 anywhere between 10 to 16 copies per cell.²¹ The gene *aph(3')-IIa*, is located on an IncI2 plasmid,
281 which conversely is a low copy number plasmid²², and is consistent with our results. The optimization

282 of future versions of this platform for specific gene targets is possible. In the case of plasmid-associated
283 genes or genes for which amplification failed, PCR conditions could be varied for optimal
284 amplification and specific gene standards could be included for absolute quantification. Further, it is
285 possible to select primers for DARTE-QM to target specific resistance classes, rather than the broad
286 array of targets demonstrated in this study.

287

288 A limitation of DARTE-QM is the presence of biased PCR amplification and associated amplicon
289 artifacts. These sequencing artifacts were observed in all samples in this study and could be
290 distinguished by the presence of a primer with an untargeted sequence. While these genes could be
291 non-specific amplification of primers targeting other biological genes, the presence of poly-A and poly-
292 T sequence patterns, like those seen in single cell amplification²³, along with their majority singleton
293 presence, suggested that they were sequencing artifacts. While these artifacts present an impediment for
294 leveraging the sequencing coverage of DARTE-QM, we found that with at least 25,000 reads per
295 sample, we could identify 90% of the ARGs present in mock-community samples. These sequencing
296 artifacts also seemed to be produced by particular primers and in samples from specific environments,
297 suggesting opportunities for optimization in future development of DARTE-QM. For instance, the
298 primers targeting vancomycin-associated ARGs produced large number of artifact reads, and no
299 vancomycin ARGs were expected in any of our samples. Similarly, many of the samples that produced
300 the highest percentage of reads as artifacts were from soils, a medium known to have PCR inhibitors²⁴.
301 In samples where there was high-quality DNA and lower diversity (e.g., mock-community samples), it
302 did not appear that the artifacts obstructed the production of true- positive reads. For screening of a
303 broad range of diverse environments, artifacts are easily filtered through target alignment and
304 classification. Future studies aimed at improving the sequencing library preparation protocols for

305 sample types or ineffective primers will continue to improve the platform based on the knowledge
306 gained.

307

308 The most beneficial aspect of DARTE-QM to improving microbiome ARG monitoring is its ability to
309 detect ARGs at costs that will allow hundreds of samples to be screened simultaneously. A current
310 challenge to antimicrobial resistance monitoring is that characterizing broad indicators are expensive,
311 and thus it is difficult to standardize studies for monitoring. DARTE-QM is a complement to existing
312 approaches to characterize ARGs. We envision an optimal system whereby the most relevant ARGs in a
313 study can be detected with less bias using metagenome sequencing, and these ARGs can subsequently
314 be targeted for numerous samples using DARTE-QM. The sequencing from DARTE-QM can then
315 provide information on the distribution of ARGs, as well as sequence variants, in a systematic fashion,
316 even if in low abundance.

317

318 DARTE-QM is the first demonstration of simultaneous library preparation and subsequent sequencing
319 of hundreds of unique gene targets from environmental DNA. Here, we demonstrated this application
320 for the characterization of ARGs and associated resistomes in environmental samples, however,
321 DARTE-QM presents the opportunity to apply this approach towards gaining sequencing information
322 for other diverse functional genes as well. This platform is particularly suited for studies in which
323 genes of interest are numerous and well-defined, and where sequencing information from these genes
324 would provide benefits to understanding biological operations (e.g., point mutations or association with
325 sequences with host information). The ability to affordably scale for numerous genes and samples
326 provides a much-needed resource for not only the field of antimicrobial resistance but for researchers
327 interested in scaling functional gene characterization. Finally, we recognize that this is the first
328 evaluation of DARTE-QM and that there are significant opportunities to further develop this approach

329 for more targeted study. Given the simultaneous amplification of primers in DARTE-QM, we expect
330 that the more specific the gene targets, the more optimized the library preparation can be for reliable
331 quantification.

332

333

334 **DATA AVAILABILITY**

335 Sequence files, sample metadata, and the genome sequence for the mock-community member
336 sequenced by the USDA facility in Ames, IA, can be found through FileShare this link

337 <https://doi.org/10.25380/iastate.14390342>

338

339 Alternatively, all metadata and mock genomes used in the study are available through the same
340 repository as the code for analysis.

341

342

343 **CODE AVAILABILITY**

344

345 All code used for processing and analysis is open-source and can be found at

346 <https://schuyler-smith.github.io/DARTE-QM/>

347

348 **ONLINE METHODS**

349

350 **Sequencing Targets and Primer Design**

351

352 Antibiotic resistance gene (ARG) targets for primer design were chosen and aggregated from two
353 sources. There were 2,472 sequences were obtained from the ResFinder database (version 3.2,
354 November, 2016)²⁵, associated with 67 antibiotic resistance families. ResFinder was selected on
355 account of its manual curation of genes associated with acquired antibiotic resistance. An additional
356 409 ARG-associated sequences chosen as well, which had previously demonstrated high prevalence in
357 animal agriculture¹⁹. To abide with the limitation of the number of allowed primers with the Illumina
358 TruSeq library preparation, later described, the conglomerate of the chosen sequences was ultimately
359 curated to representative sequences that targeted genes deemed of most interest to antibiotic resistance
360 in agriculture. A single 300 bp synthetic oligonucleotide sequence was designed for use as a reference
361 (reference target gene in Supp. Table 1). The synthetic oligonucleotide was designed with no biological
362 context to ensure that it would not interfere with any ARG detection, save for appropriate restriction
363 sites that were added to allow for insertion into a pUC19 cloning vector. The sequence was compared
364 to the entirety of the NCBI Genbank database and was confirmed to share no significant similarity to
365 any existing records. Lastly, we included 25 sequences based on those used by the Earth Microbiome
366 Project²⁶ to target the V4 variable region of the 16S rRNA gene.

367

368 The goal of primer design was to target the maximum number of our chosen sequences, with the
369 highest specificity, staying within the set limit of 1,536 primers for the library preparation. Primers
370 were designed using the Ribosomal Database Project's EcoFunPrimer software:²⁷ product minimum
371 length = 220, product maximum length = 330, Oligo minimum size = 22, oligo maximum size = 30,
372 maximum mismatch = 0, temperature minimum = 55, temperature maximum = 63, hair-pin max = 24,
373 homo-max = 35, assaymax = 30, degenmax = 6, noTEendfileter = T, nopoly3GCfilter = T, polypyrunfilter
374 = 4, GCfilter min = 0.15 GCfilter max = 0.8. This produced 1,340 primers (Supp. Table 1) to target the
375 ARG associated sequences, which accounted for 2,184 sequences (88.3%) from those selected (Supp.

376 Table 2). Two primers were created for the synthetic oligonucleotide, and 30 were included for
377 targeting all degeneracies of the 25 16S rRNA sequences. In total, DARTE-QM used 1,372 primers
378 (668 forward-primers, 704 reverse-primers) for 796 primer pairs to be used with Illumina's TruSeq
379 Custom Amplicon Low Input library preparation. These primers targeted representative sequences of
380 all 67 antibiotic resistant families and 662 ARGs.

381

382 **Library Prep**

383 Oligonucleotide primers were created in Illumina Design Studio and ordered through Illumina (Supp.
384 Table 1). Paired-end libraries for each sample were prepared using the TruSeq Custom Amplicon Low
385 Input Kit (Illumina) according to the manufacturer's instructions. This kit allows generation of up to
386 1536 amplicon targets over 96 samples. All DNA was diluted to 10 ng/uL during library preparation, or
387 prepared with no dilution where concentrations were less than 10 ng/uL. An Agilent High Sensitivity
388 D1000 ScreenTape System (Agilent Technologies) was used for measuring DNA concentration of pre-
389 pared libraries. For sequencing, the MiSeq Reagent Kit v2 (300-cycles) (Illumina) reagents were used
390 with the MiSeq sequencing platform.

391 **Samples**

392

393 **Mock-community**

394 A mock-community composed of 20 cultured isolates²⁸ was created for purpose of assessing the
395 effectiveness of DARTE-QM. Nineteen of the genomes were available from the NCBI GenBank, and a
396 single genome was sequenced at the USDA Animal Research (Ames, Iowa) (Supp. Dataset 1). The
397 ARGs found within the genomes were annotated using ResFam and also the Comprehensive Antibiotic

398 Resistance Database (CARD, version 2.0.1)²⁹. We included 6 mock-community samples sequenced in
399 triple replicates with 0, 0.0025, 0.009, 0.025, 0.12, 0.25 ng of the synthetic oligonucleotide reference.

400

401 To evaluate the practical implementation of DARTE-QM using environmental samples, we used 19
402 environmental samples originating from intrinsic and manure-amended soils, swine manures, effluent
403 from manure-amended soils, and swine fecal samples that passed quality filters. Samples were selected
404 from two previously published studies. In the first study, laboratory soil columns and rainfall
405 simulations were used to evaluate the influence of swine manure amendment on soils and effluent¹⁹
406 (Supp. Table 4). In the second study, fecal samples from swine with varying antibiotic usage and routes
407 of administration were used³⁰. Samples from a subsequent laboratory soil column experiment designed
408 to evaluate the influence of either swine or beef manure on soils and effluent were also included.
409 Metagenomes were available for 14 samples (NCBI SRA database Bioproject PRJNA533779) were
410 used for comparisons to DARTE-QM results.

411

412 Data Analysis

413

414 All analysis was done in the statistical language R, unless otherwise stated. DARTE-QM sequences
415 were quality checked using FastQC (v0.11.9)³¹ (Figure 1). Reads were demultiplexed by primer, which
416 were identified and removed using Cutadapt (v2.10)³² with an error tolerance of 0.1 and a phred-score
417 quality threshold of 20³². High-quality paired-end reads were merged using PEAR (v0.9.8)³³, requiring
418 a minimum overlap of 10 bp. Merged reads were aligned against our database of targeted sequences
419 using BLAST (v2.10)³⁴. Successful alignment required a minimum of 90 bp and 98% similarity. For
420 paired-end reads that were not able to be merged, each was aligned to the target-database individually.
421 If both reads aligned to the same target, the read with the longest alignment was selected as the repre-

422 sentative sequence. We defined a successful amplification as a read for which a primer sequence was
423 present, and the amplified sequence aligned to the primer's intended target with at least 90 bp length
424 and at least 95% identity. Reads identified as having 16S rRNA primers were classified using the RDP
425 Classifier³⁵ with default parameters, and then unpaired reads selected in the same manner as for ARGs.

426

427 Each read was classified as either true positive (TP), false positive (FP), false negative (FN), or true
428 negative (TN). This classification was based on alignment to the ARG reference database, where reads
429 were deemed to be TP when both the intended primer target and read sequence aligned to the same
430 gene; FP when the primer target and the read sequence did not agree; FN when no primer was found
431 but the read sequence was able to be aligned to a reference ARG; and TN reads were assigned as all
432 reads within a sample assigned as TP outside of the primer in question, i.e., all reads that were correctly
433 identified as not being the targeted read. Success for DARTE-QM was evaluated on three metrics: sen-
434 sitivity ($TP/[TP + FN]$), specificity ($TN/[TN + FP]$), and accuracy ($[TP + TN]/[TP + FN + TN + FP]$)
435 for each gene target (i.e., primer) and each sample.

436

437 The ability of DARTE-QM to quantify ARG presence was tested by comparing observed counts of the
438 synthetic oligonucleotide to the expected concentrations. Samples were normalized by rarefying to a
439 sequence count of 5,000. Samples with a sequence count less than 5,000 were discarded. Alpha
440 diversity, richness, of ARGs was calculated using Shannon's index. Principal coordinate analysis was
441 conducted to evaluate the variations of resistome profile in samples. Based on the relative abundance of
442 ARGs in each sample, Bray-Curtis distances were calculated for each pair of samples, and the first two
443 components of the eigenvalue decomposition were plotted. Permutational multivariate analysis of
444 variance (PERMANOVA) was used to identify the significant factors (e.g., experiments, source-

445 matrices) which contributed to the observed resistome variation. Cluster analysis was performed using
446 k-means.

447

448 **Variant Analysis**

449

450 To evaluate the presence of gene sequence variants, the observations of variants were estimated for
451 sequences associated with the *erm35* and *tetM* genes. The forward reads of sequences which aligned to
452 the DARTE-QM gene targets were clustered at 97% sequence similarity with CD-HIT (v4.6.7)³⁶.
453 Clusters containing greater than ten sequences were considered in our results, with representative
454 sequences for each cluster determined by CD-HIT. Alignment was performed and visualized with
455 JalView using ClustalW (v2.11.1.3)³⁷.

456

457

458 **REFERENCES**

459

- 460 1. PCAST. National action plan for combatting antibiotic-resistant bacteria. *Washington, DC White*
461 *House* (2015).
- 462 2. WHO. Global action plan on antimicrobial resistance. *World Heal. Organ.* 1–28 (2017).
- 463 3. Nelson, R. E. *et al.* National Estimates of Healthcare Costs Associated With Multidrug-Resistant
464 Bacterial Infections Among Hospitalized Patients in the United States. *Clin. Infect. Dis.* **72**, S17–
465 S26 (2021).
- 466 4. Nowakiewicz, A. *et al.* A significant number of multi-drug resistant *Enterococcus faecalis* in
467 wildlife animals; long-term consequences and new or known reservoirs of resistance? *Sci. Total*
468 *Environ.* **705**, 135830 (2020).

469 5. Werner, G. *et al.* Antibiotic resistant enterococci—tales of a drug resistance gene trafficker. *Int.*
470 *J. Med. Microbiol.* **303**, 360–379 (2013).

471 6. Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome:
472 Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res.*
473 *Commun.* **469**, 967–977 (2016).

474 7. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics,
475 from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

476 8. Wommack, K. E., Bhavsar, J. & Ravel, J. Metagenomics: read length matters. *Appl. Environ.*
477 *Microbiol.* **74**, 1453–1463 (2008).

478 9. Guitor, A. K. *et al.* Capturing the resistome: a targeted capture method to reveal antibiotic
479 resistance determinants in metagenomes. *Antimicrob. Agents Chemother.* **64**, (2019).

480 10. Noyes, N. R. *et al.* Enrichment allows identification of diverse, rare elements in metagenomic
481 resistome-virulome sequencing. *Microbiome* **5**, 1–13 (2017).

482 11. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences
483 per sample. *Proc. Natl. Acad. Sci.* **108**, 4516–4522 (2011).

484 12. Beltrame, L. *et al.* Profiling cancer gene mutations in longitudinal epithelial ovarian cancer
485 biopsies by targeted next-generation sequencing: a retrospective study. *Ann. Oncol.* **26**, 1363–
486 1371 (2015).

487 13. Chen, R., Im, H. & Snyder, M. Whole-exome enrichment with the illumina truseq exome
488 enrichment platform. *Cold Spring Harb. Protoc.* **2015**, pdb--prot084863 (2015).

489 14. Williams, E. L. *et al.* Performance evaluation of Sanger sequencing for the diagnosis of primary
490 hyperoxaluria and comparison with targeted next generation sequencing. *Mol. Genet. \&*
491 *genomic Med.* **3**, 69–78 (2015).

492 15. Wong, S. Q. *et al.* Assessing the clinical value of targeted massively parallel sequencing in a

493 longitudinal, prospective population-based study of cancer patients. *Br. J. Cancer* **112**, 1411–
494 1420 (2015).

495 16. Bürgmann, H., Pesaro, M., Widmer, F. & Zeyer, J. A strategy for optimizing quality and quantity
496 of DNA extracted from soil. *J. Microbiol. Methods* **45**, 7–20 (2001).

497 17. Flekna, G., Schneeweiss, W., Smulders, F. J. M., Wagner, M. & Hein, I. Real-time PCR method
498 with statistical analysis to compare the potential of DNA isolation methods to remove PCR
499 inhibitors from samples for diagnostic PCR. *Mol. Cell. Probes* **21**, 282–287 (2007).

500 18. Tebbe, C. C. & Vahjen, W. Interference of humic acids and DNA extracted directly from soil in
501 detection and transformation of recombinant DNA from bacteria and a yeast. *Appl. Environ.*
502 *Microbiol.* **59**, 2657–2665 (1993).

503 19. Smith, S. D. *et al.* Investigating the dispersal of antibiotic resistance associated genes from
504 manure application to soil and drainage waters in simulated agricultural farmland systems. *PLoS*
505 *One* **14**, 1–14 (2019).

506 20. Holman, D. B., Brunelle, B. W., Trachsel, J. & Allen, H. K. Meta-analysis to define a core
507 microbiota in the swine gut. *MSystems* **2**, e00004--17 (2017).

508 21. Rawlings, D. E. & Tietze, E. Comparative biology of IncQ and IncQ-like plasmids. *Microbiol.*
509 *Mol. Biol. Rev.* **65**, 481–496 (2001).

510 22. Rozwandowicz, M. *et al.* Plasmids carrying antimicrobial resistance genes in
511 Enterobacteriaceae. *J. Antimicrob. Chemother.* **73**, 1121–1137 (2018).

512 23. Blainey, P. C. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol.*
513 *Rev.* **37**, 407–427 (2013).

514 24. Lim, H. J., Choi, J.-H. & Son, A. Necessity of purification during bacterial DNA extraction with
515 environmental soils. *Environ. Health Toxicol.* **32**, (2017).

516 25. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob.*

517 26. *Chemoth. 67*, 2640–2644 (2012).

518 26. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and
519 aspirations. *BMC Biol. 12*, 69 (2014).

520 27. Cole, J. R. *et al.* Ribosomal Database Project: Data and tools for high throughput rRNA analysis.
521 *Nucleic Acids Res. 42*, 633–642 (2014).

522 28. Allen, H. K. *et al.* Pipeline for amplifying and analyzing amplicons of the V1--V3 region of the
523 16S rRNA gene. *BMC Res. Notes 9*, 380 (2016).

524 29. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive
525 antibiotic resistance database. *Nucleic Acids Res. 48*, D517--D525 (2020).

526 30. Ricker, N. *et al.* Toward antibiotic stewardship: Route of antibiotic administration impacts the
527 microbiota and resistance gene diversity in swine feces. *Front. Vet. Sci. 7*, (2020).

528 31. Andrews, S. & others. FastQC: a quality control tool for high throughput sequence data. (2010).

529 32. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
530 *EMBnet. J. 17*, 10–12 (2011).

531 33. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End
532 reAd mergeR. *Bioinformatics 30*, 614–620 (2014).

533 34. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search
534 tool. *J. Mol. Biol. 215*, 403–410 (1990).

535 35. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid
536 assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol. 73*,
537 5261–5267 (2007).

538 36. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation
539 sequencing data. *Bioinformatics 28*, 3150–3152 (2012).

540 37. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version

541 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191
542 (2009).

543

544

545 **ACKNOWLEDGEMENTS**

546 This project was supported (or partially supported) by AFRI food safety grant no. 2016-68003-24604
547 from the USDA National Institute of Food and Agriculture. We thank Jennifer Jones and Kathy Mou at
548 the *ARS-USDA National Animal Disease Center* for their help with library preparation. We thank Jared
549 Shelerud at Illumina for his help with the TruSeq Custom Amplicon platform.

550

551 **AUTHOR CONTRIBUTIONS**

552 A.H., H.A., M.S., F.Y., N.R., and J.C. were designed the project; A.H. H.A., M.S., and S.H.-L. were
553 involved in funding-acquisition; S.S. analyzed the data and wrote the manuscript with assistance from
554 A.H. and N.R.

555

556 **COMPETING INTERESTS**

557 The authors declare no competing interests.

558

559

560 **ADDITIONAL INFORMATION**

561 Supplemental information

562

563 **CORRESPONDENCE**

564 Correspondence to Adina Howe

565

566 PEER REVIEW INFORMATION

567

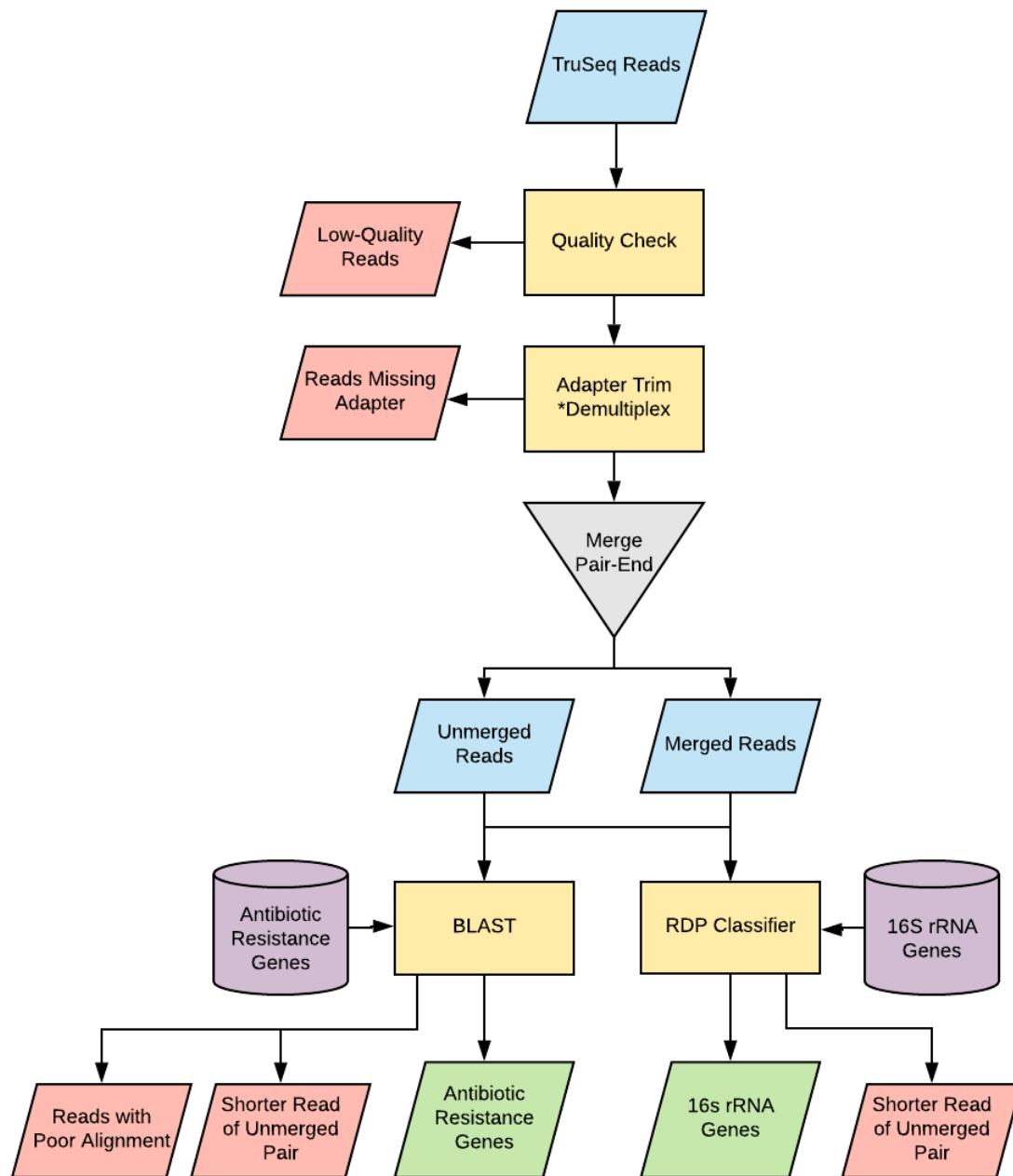
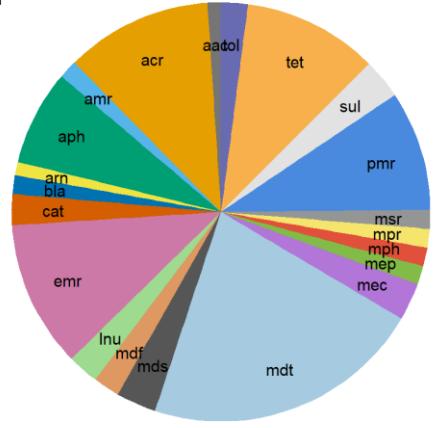
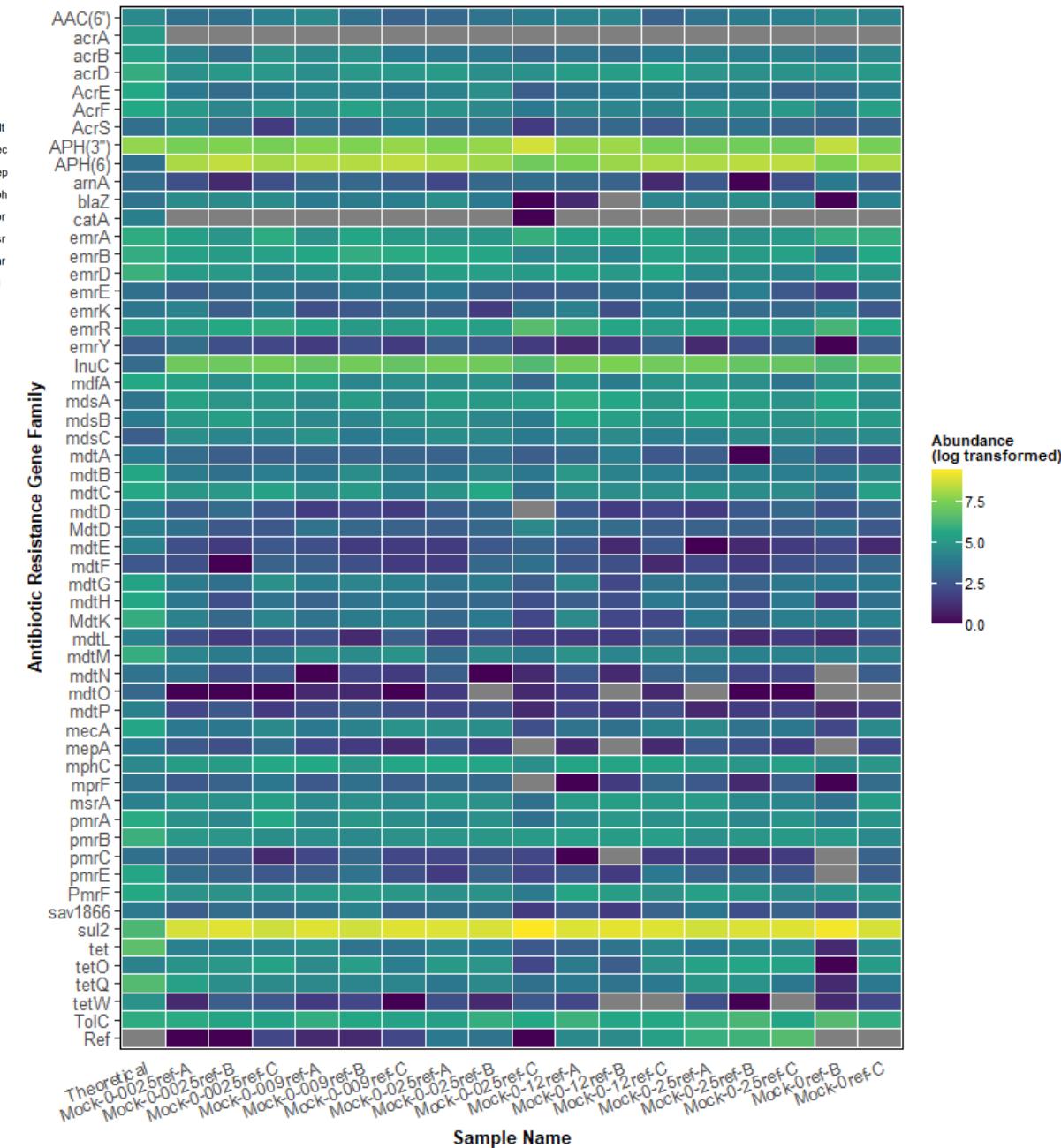


Figure 1. Summary of DARTE-QM read processing pipeline. Data-boxes color blued represent reads kept within the pipeline, red boxes were discarded reads, and green are the finalized reads for analysis. Reads were filtered by quality-score and demultiplexed by the presence of primer sequences. To classify ARGs, both merged and unmerged reads were required to align to known genes in ResFam and CARD ARG reference databases. In the case of unmerged reads, if both the forward and the reverse read aligned to the same target, the shorter alignment from the pair was discarded.

a**b**

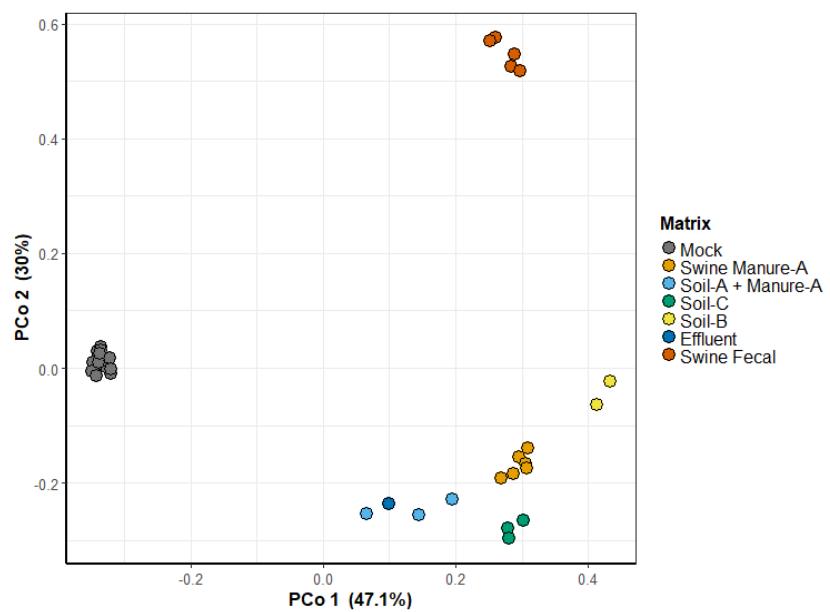
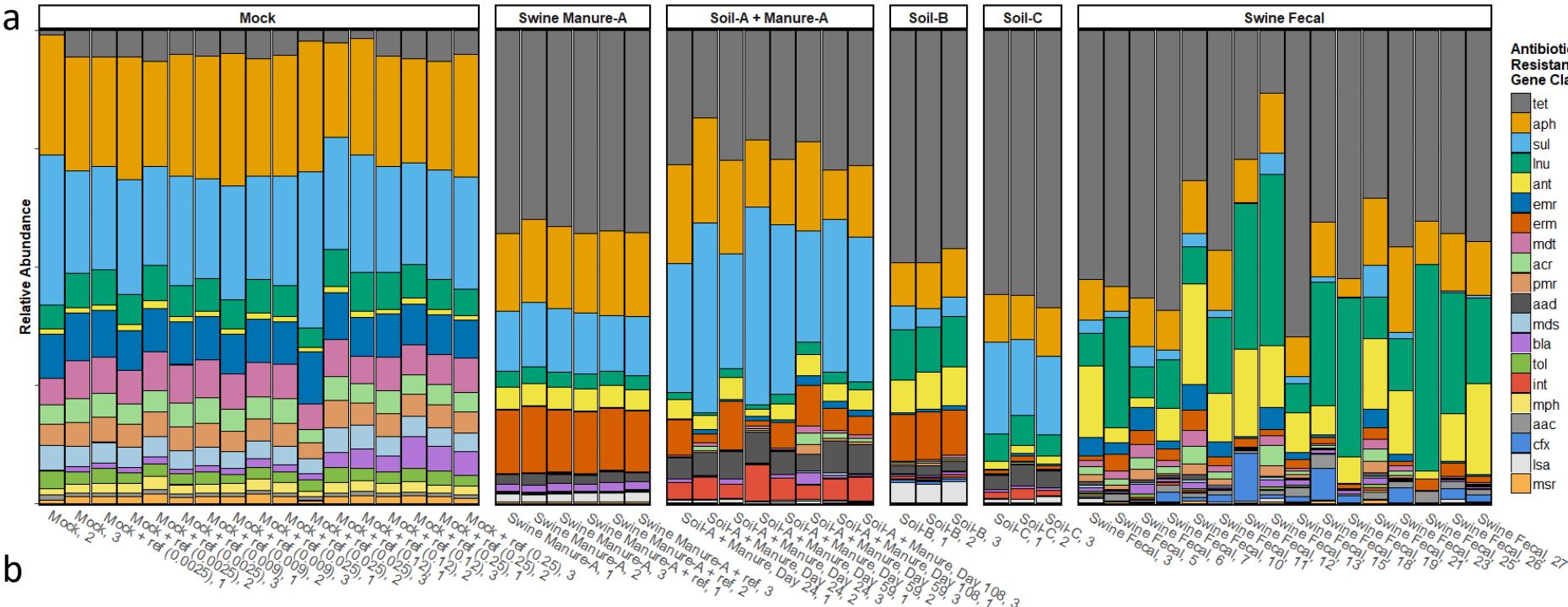
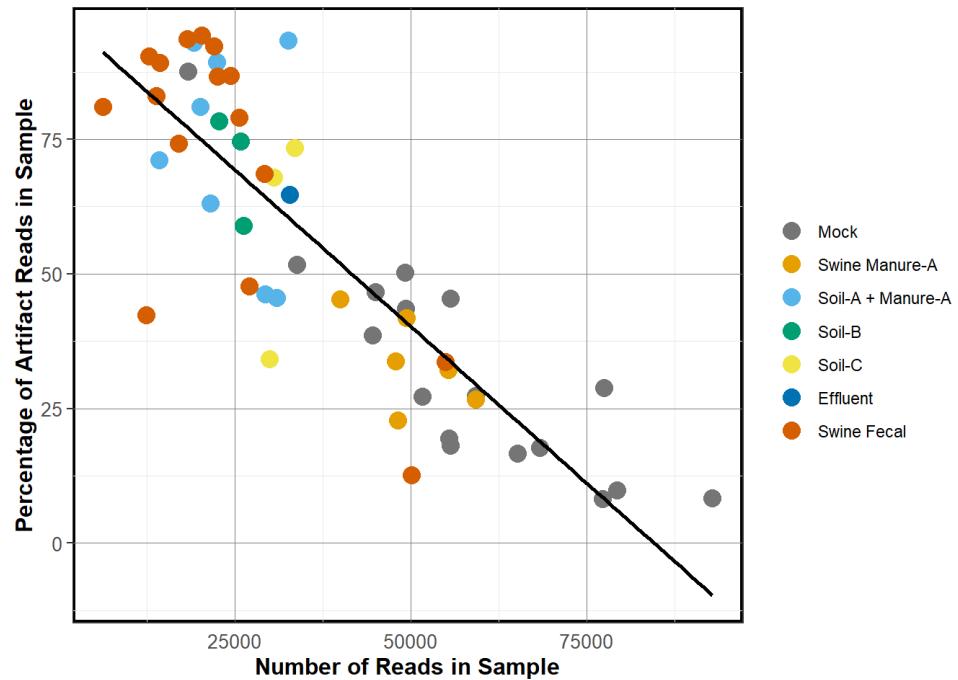
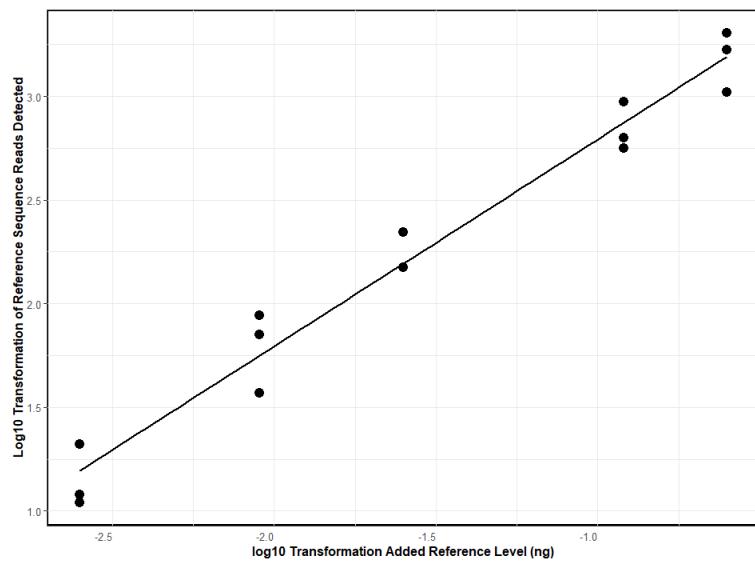
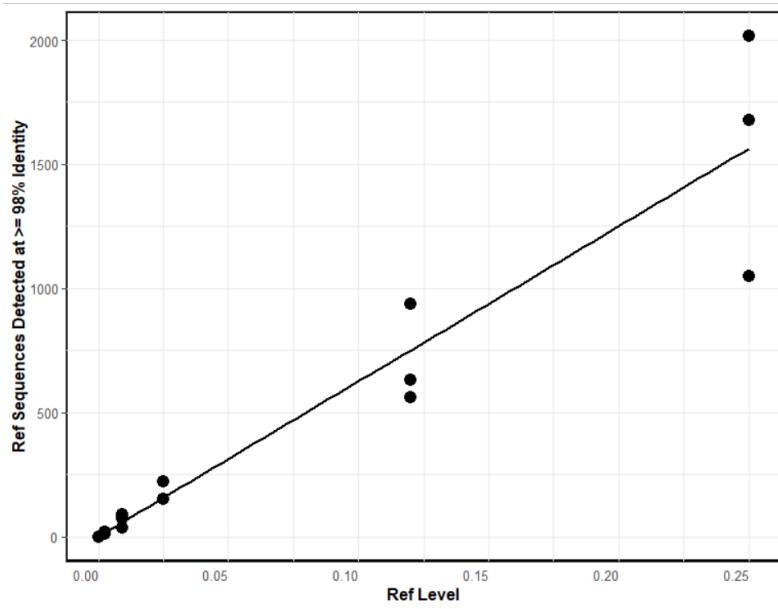


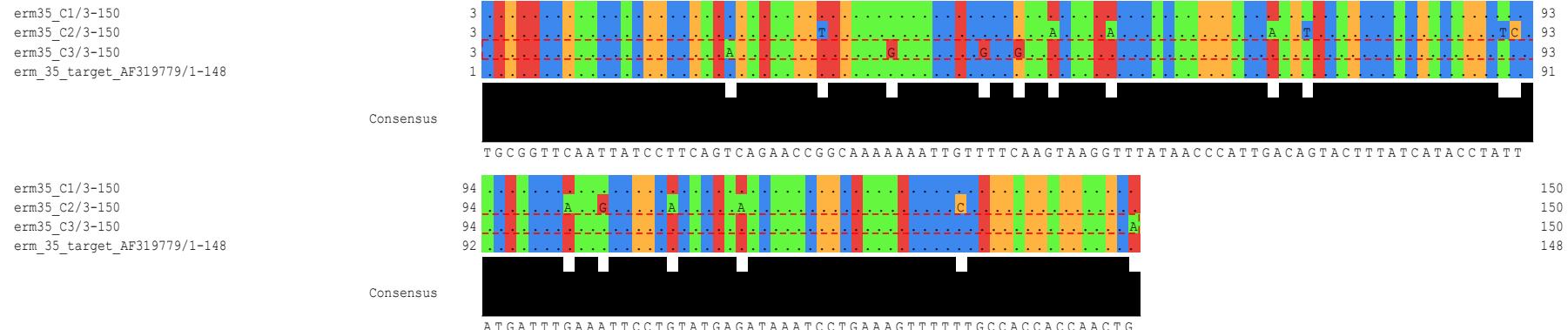
Figure 3. ARG profiles by source matrix. a) Relative abundance of ARG classes identified for all mock community, swine fecal, soil, swine manure, and manure-treated soils. b) Principal coordinate analysis based on Bray-Curtis distances of for resistomes, for samples passing all QC-filtering.



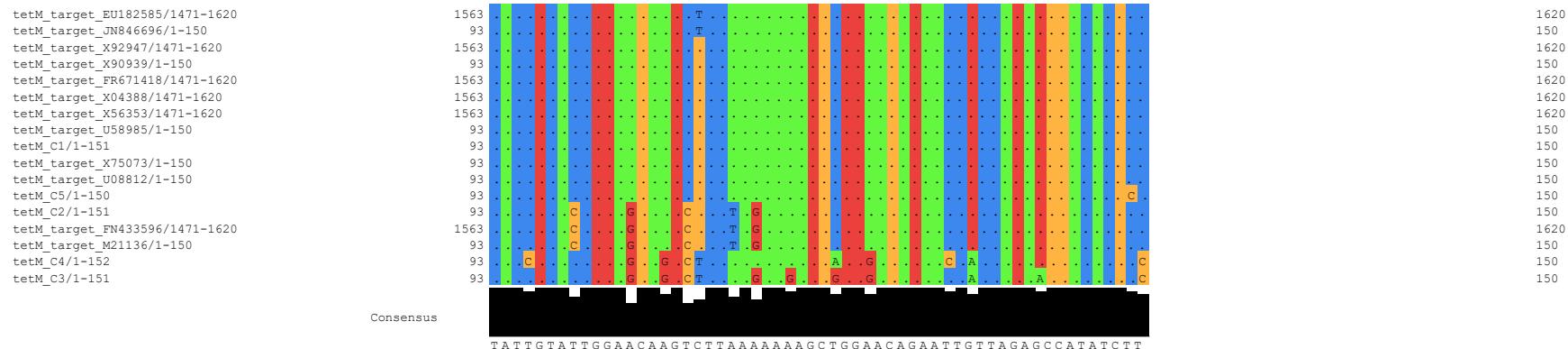
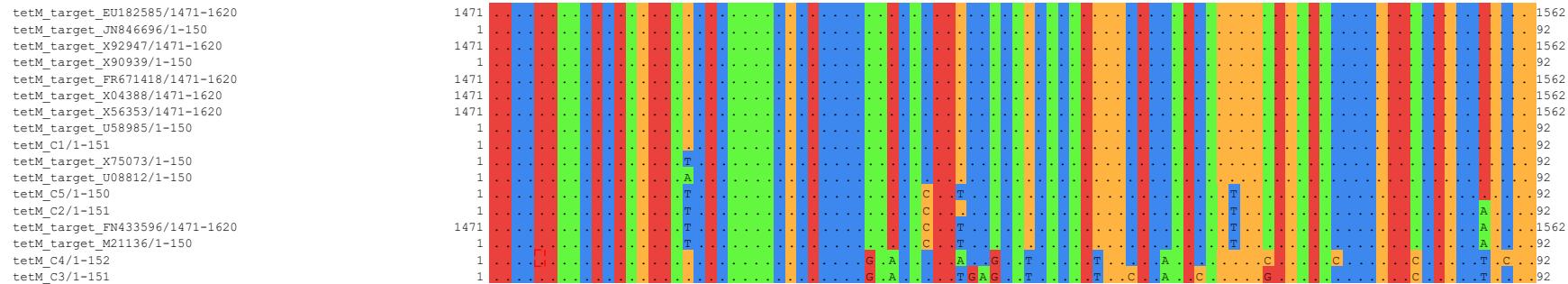
Supp Figure 1. Linear correlation of the percentage of artifacts reads present in a sample to the total number of reads in the sample. Reads were defined as sequencing artifacts if a primer was located on the 5' end of the sequences and the read did not align to any of reference ARGs or any other location in the mock-community genomes. The percentage of sequencing artifacts observed was higher for environmental samples relative to mock community samples and was also inversely correlated ($R^2 = 0.68$) to the number of reads in a sample.



Supp Figure 2. Linear correlation between the concentration of the reference sequence added to mock community samples and the number of reads which aligned to the reference sequence. The linear model found there to be a strong correlation ($R^2 = 0.91$), indicating DARTE-QM was sensitive to DNA quantity.



Supp Figure 3. a) Alignment of gene targets and sequences identified by DARTE-QM gene target for a) *erm35* and [continued on next slide]



Supp Figure 3. b) *tetM* genes. Sequences are representative sequences identified for clusters of reads by 97% sequence identity (see also Supp. Table 7). The presence of gene variants are shown in the corresponding three clusters for *erm35* (*erm35_C1-C3*) and five clusters for *tetM* (*tetM_C1-C5*). Genes targeted by DARTE-QM are also shown.