

1 Supporting Information 1

2

3 Metagenomic analysis of antimicrobial resistance genes in wastewaters in Benin and Burkina Faso indicates a  
4 serious health risk from untreated hospital wastewaters in low-income countries

5

6 Melina A. Markkanen<sup>1\*</sup>, Kaisa Haukka<sup>1,6</sup>, Katariina M. M. Pärnänen<sup>1</sup>, Victorien Tamegnon Dougnon<sup>2</sup>, Isidore  
7 Juste O. Bonkougou<sup>3</sup>, Zakaria Garba<sup>4</sup>, Halidou Tinto<sup>4</sup>, Anniina Sarekoski<sup>1</sup>, Antti Karkman<sup>1</sup>, Anu Kantele<sup>5,6†</sup>,  
8 Marko P. J. Virta<sup>1\*†</sup>

9

10 <sup>1</sup>Department of Microbiology, University of Helsinki, Finland

11 <sup>2</sup>Research Unit in Applied Microbiology and Pharmacology of Natural Substances, Polytechnic School of  
12 Abomey-Calavi, University of Abomey-Calavi, Benin

13 <sup>3</sup>Department of Biochemistry and Microbiology, University Joseph KI-ZERBO, Ouagadougou, Burkina Faso

14 <sup>4</sup>Clinical Research Unit Nanoro, Institute for Research in Health Sciences, National Center for Scientific and  
15 Technological Research, Burkina Faso

16 <sup>5</sup>Meilahti Infectious Diseases and Vaccine Research Center, MeiVac, Department of Infectious Diseases,  
17 University of Helsinki and Helsinki University Hospital, Helsinki, Finland

18 <sup>6</sup>Human Microbiome Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland

19

20 \*Marko P. J. Virta, [marko.virta@helsinki.fi](mailto:marko.virta@helsinki.fi)

21 \*Melina A. Markkanen, [melina.markkanen@helsinki.fi](mailto:melina.markkanen@helsinki.fi)

22 † Equal contribution

23

24 Table of Contents:

25 **Supporting Information for Bioinformatic Analyses**.....2  
26 **Supporting Information for Statistical Analyses**.....3  
27 **Figure S1.** Sample collection in Benin and Burkina Faso.....4  
28 **Figure S2.** Examples of sampling sites.....4  
29 **Figure S3.** PCoA plots for resistomes, taxonomical compositions, and mobilomes.....6  
30 **Table S1.** Significance levels for ordination plots.....7  
31 **Figure S4.** Unique and shared ARGs between hospital wastewaters in Benin, Burkina Faso, and Finland....7  
32 **Figure S5.** Relative abundance of clinically relevant species.....8  
33 **Figure S6.** Multidrug-resistance fragments putatively encoded by class 1 integron gene cassettes.....8  
34 **Figure S7.** Contigs harboring plasmid-mediated *mcr-5.1* colistin resistance.....9  
35 **References** .....10

36

37 Pages: 11

38 Figures: 7

39 Tables: 1

40

## 41 **Supporting Information for Bioinformatic Analyses**

42

43 The quality control steps included in the workflow were performed using FastQC<sup>1</sup> and MultiQC<sup>2</sup>, with adapter  
44 and low-quality read removal using Cutadapt<sup>3</sup>. Nucleotide sequence reads were mapped using Bowtie2<sup>4</sup> against  
45 the ResFinder database (version 3.2, downloaded on 2020-06-28)<sup>5</sup>. The reads were sorted and filtered using  
46 SAMtools<sup>48</sup> such that the reads mapping as pair or alone were calculated as a single count. Mobile genetic  
47 elements were identified by mapping the reads against the MobileGeneticElementDatabase  
48 (<https://github.com/KatariinaParnanen/MobileGeneticElementDatabase>, downloaded on 2020-06-28) consisting  
49 of 2714 unique MGE sequences including transposases; integrons of classes 1, 2, and 3; the integron associated  
50 disinfectant resistance gene *qacEA*; and IS and ISCR type insertion sequences<sup>6</sup>. Class 1 integrons were studied  
51 using mapping results obtained for the *intI* gene, which encodes an integron integrase (*IntI*). Taxonomic profiling  
52 was performed using both Metaphlan3<sup>7</sup> and Metaxa2<sup>8</sup>. The counts for bacterial 16S rRNA from Metaxa2 were  
53 used for normalization of ARG and MGE counts. The lengths of the functional genes and the 16S rRNA SSU  
54 gene were taken into account in the normalization.

55

56 The accuracy of the normalization method was validated by comparing two distinct methods using the 16S rRNA  
57 and *rpoB* housekeeping genes. The *rpoB* counts were obtained using the translated sequence reads described above  
58 as input for hmmsearch<sup>9</sup> against a *rpoB* specific HMM profile (pf04563) from the Pfam-A database<sup>10</sup>. Hits for  
59 forward reads (R1) were counted. The correlation between the ARGs normalized to both 16S rRNA and *rpoB* was  
60 significant ( $R^2 = 0.61$ ,  $p = 4.8 \times 10^{-8}$ ), and we proceeded to use 16S rRNA genes for normalization.

61

62 To study plasmid-mediated colistin resistance genes (*mcr* genes) and their genetic environments for their  
63 theoretical mobility potential, a subset of the samples was assembled using MEGAHIT (v1.2.8)<sup>11</sup>, including the  
64 samples BH02, BH48, BFH19, BFH26, BFH41, FH1, and FH9. The minimum contig length was set to 1,000 nt.  
65 Contigs harboring plasmid-mediated colistin resistance genes were identified using a BLAST<sup>12</sup> search against *mcr*  
66 genes in the ResFinder database (n=54). Similarly, contigs containing class 1 integrons were identified and  
67 screened for putative ARGs located in the same contig.

68

69 Visualizing the genetic environments of contigs harboring *mcr-5.1* colistin ARGs was performed using the anvi'o  
70 (v7)<sup>13</sup> platform tool 'anvi-interactive'. For this, single profiles were formed for the samples BH02, BH48, BFH26,  
71 BFH41, and FH9 with the option '-cluster-contigs'. Then, a BLAST nr search of the contigs against the ResFinder  
72 and MobileGeneticElementDatabase was performed. The resulting annotations were imported into the individual  
73 sample profiles containing the contigs database and open reading frames predicted by Prodigal (v2.6.3)<sup>14</sup> with the  
74 command 'anvi-import-functions'. The other predicted genes, sharing the same genetic environment with the *mcr-*  
75 *5.1* gene, were identified by inspecting the resulting annotations from the BLAST nr search on the web BLAST  
76 'Graphics' view.

77

78 In order to simplify the presentation of the relative sum abundance results of the mobile colistin resistance genes,  
79 the *mcr* genes found in ResFinder database<sup>5</sup> were clustered based on 90 % similarity in their sequence identity  
80 using CD-HIT (v4.8.1)<sup>15</sup>.

81

82 To visualize the co-localization of multiple ARGs with Bandage<sup>16</sup>, the assembled contigs were converted into  
83 assembly graphs using the MEGAHIT command 'megahit\_toolkit contig2fastg'. The Bandage program was run  
84 with a BLAST<sup>12</sup> query of a custom database consisting of ARGs present in the ResFinder database and the integron  
85 sequence-related genes *qacEA* and *intI1*.

86

### 87 **Supporting Information for Statistical Analyses**

88

89 Statistical analyses were performed in R (v3.6.3), and the results were visualized using ggplot2 (v3.3.3)<sup>17</sup>,  
90 pheatmap (v 1.0.12)<sup>18</sup>, and patchwork (v1.1.0.9000)<sup>19</sup>. Vector maps were drawn using the package rnaturalearth  
91 (v0.1.0)<sup>20</sup>.

92

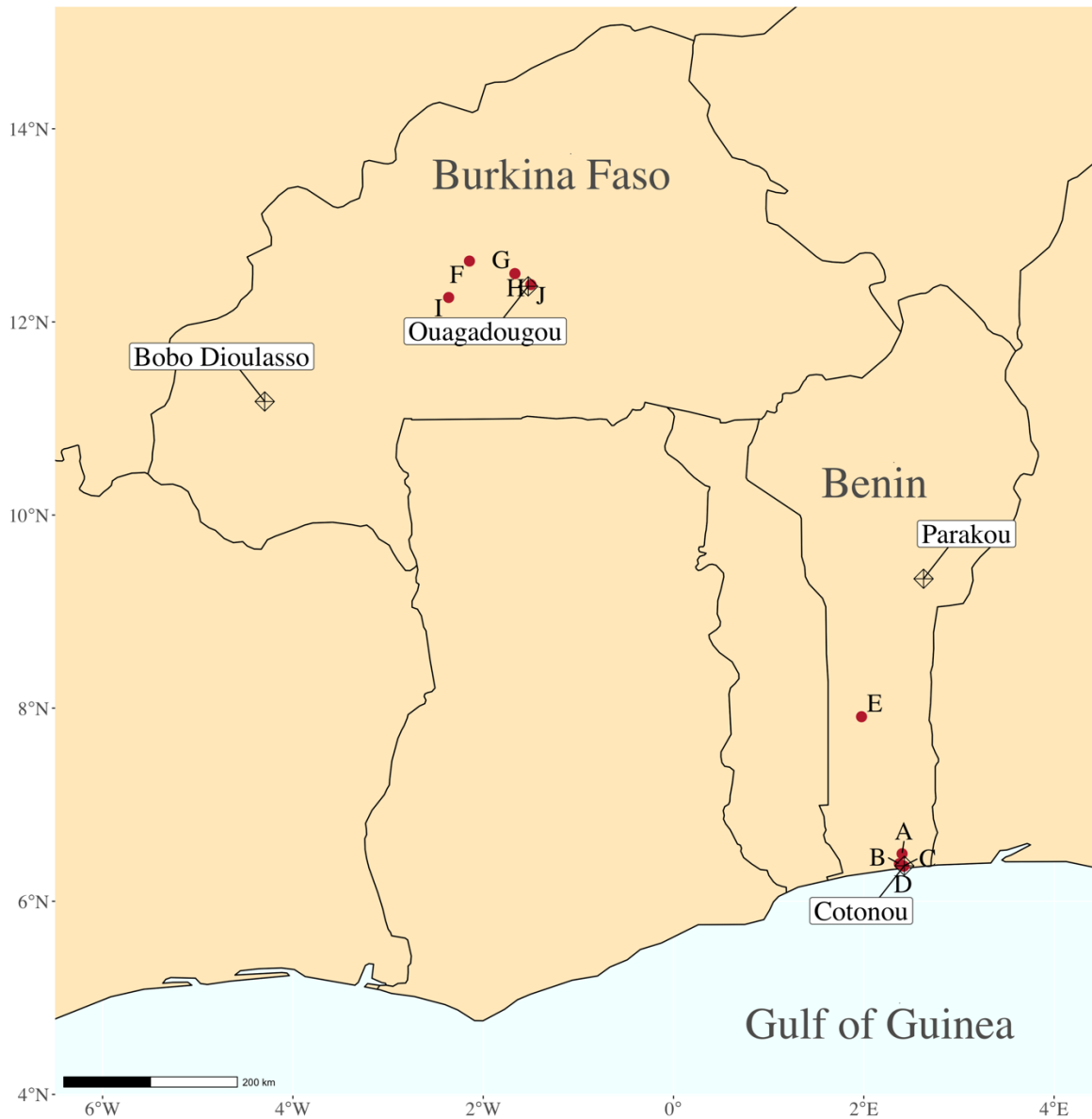
93 All statistical analyses described below were performed for all 67 HWW samples described in the 'Sample details'  
94 section and Supporting Information Table S2B. The various wards, clinics, and other units typically had their own  
95 septic tanks or sumps in the Beninese and Burkinabe hospitals. The samples were designated accordingly as  
96 'surgery', 'maternity', 'pediatry', or 'other'. In addition, samples from septic tanks or sumps receiving wastewater  
97 from different sources were designated as 'other'.

98

99 The significance of the distinct clustering between the HWWs from different countries drawn using Phyloseq  
100 package (v1.30.0)<sup>21</sup> was tested using pair-wise adonis in the package vegan (v2.5-7)<sup>22</sup> with equal sample groups.  
101 For this, eight samples per country were selected randomly among the HWWs. Unique and shared ARGs were  
102 visualized with the function 'draw.triple.venn' in the package VennDiagram (v1.6.20)<sup>23</sup>.

103

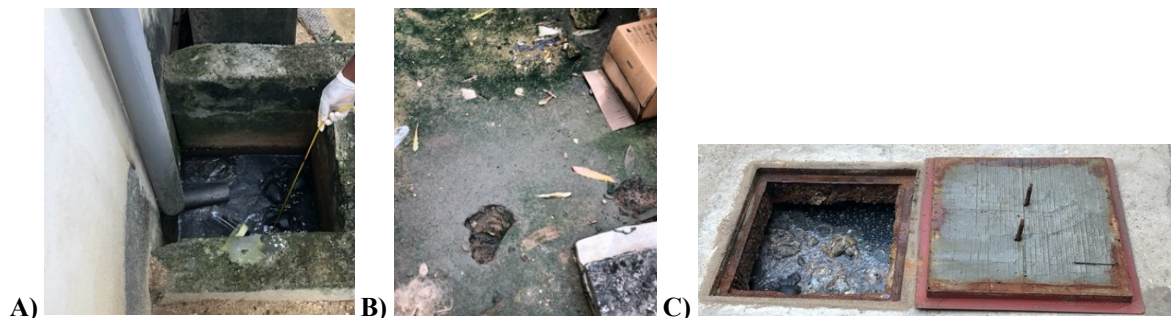
104 DESeq2 (v1.26.0)<sup>24</sup> was used to study the differentially abundant ARGs between the HWWs from the three  
105 different countries. ARGs filtered using a variance threshold of 50 were screened for significantly ( $p < 0.05$ )  
106 different relative abundances represented by log-2-fold changes across the HWWs of the three countries in a pair-  
107 wise manner (Benin-Finland, Burkina Faso-Finland). The DESeq2 algorithm is designed for unnormalized count  
108 data from gene expression studies. Thus, to take into account the amount of bacterial DNA in the samples (counts  
109 of detected 16S rRNA), the relative abundances were multiplied by  $10^5$  and a pseudo count of one was added to  
110 avoid zero values unsuitable for integer transformation. A similar analysis using DESeq2 was performed to study  
111 differentially abundant taxa found in the HWWs from Benin or Burkina Faso versus Finland. The relative  
112 abundances obtained from Metaphlan3<sup>7</sup> were multiplied with the corresponding 16S rRNA counts for each sample  
113 from the Metaxa2<sup>8</sup> output. Analyses for differentially abundant genera and species were conducted separately.



114  
115  
116  
117  
118  
119  
120  
121

**Figure S1. Sample collection in Benin and Burkina Faso.** Sample collection sites are denoted with red dots on the map. The climate in the coastal region of Benin is hot and humid. The average daytime temperature in Cotonou varies between 28–32 °C and humidity between 77–84 %, with an annual rainfall of 1,308 mm. Burkina Faso is characterized by a dry tropical climate. The average daytime temperature in Ouagadougou varies between 31–39 °C, and the annual rainfall is 745 mm. During the sampling period in November/December, it typically rains only a little in Cotonou and not at all in Ouagadougou.

122



123

D)



E)



F)



124

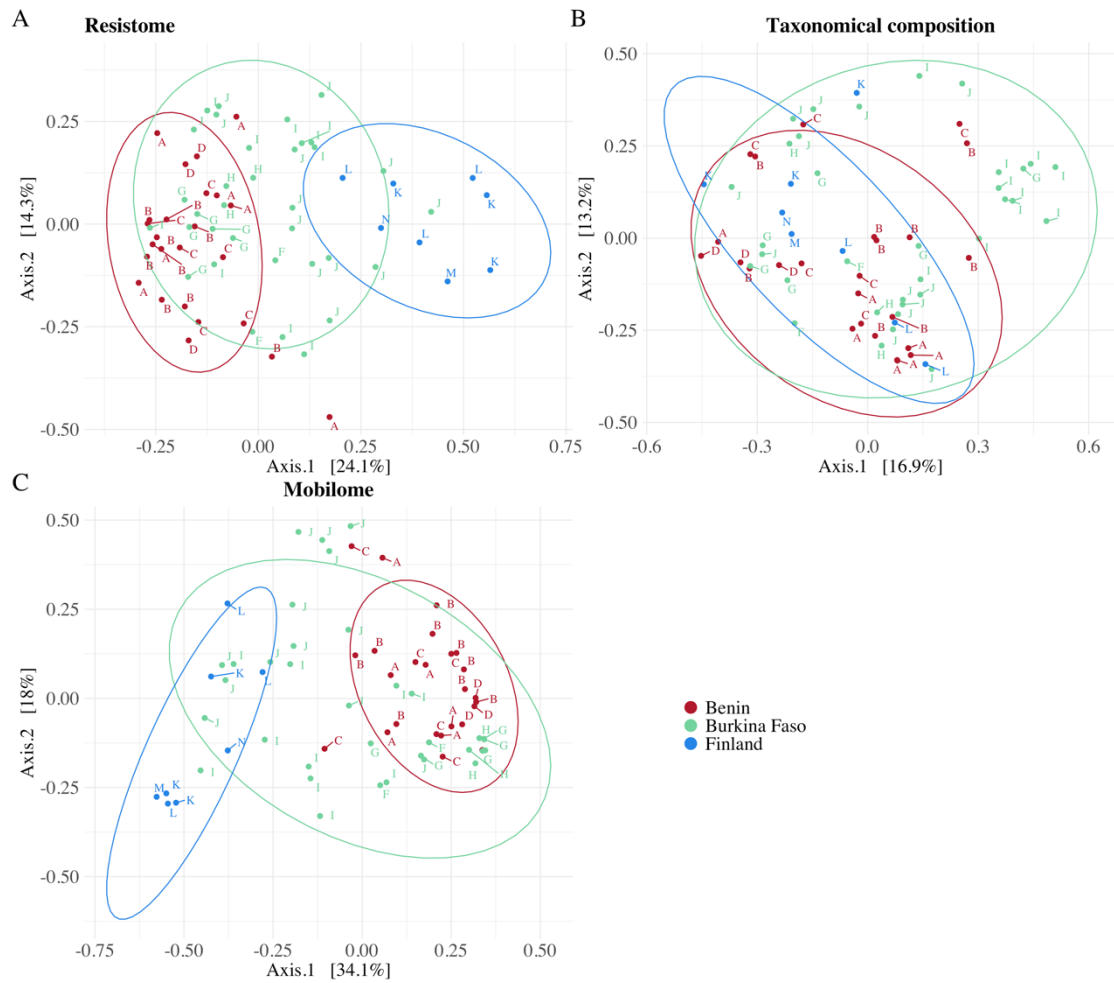
G)



H)



125 **Figure S2. Examples of sampling sites.** The samples collected in Benin included **A)** HWW from an open cesspit  
126 next to the wall at the hospital B (sample BH29); **B)** moist soil between two septic tanks (samples BH31 and  
127 BH34) at hospital B; **C)** the central septic tank collecting wastewater from different sections of hospital C (sample  
128 BH44); **D)** water puddle surrounding the surgery room septic tank at hospital C (sample BH48); **E)** water from a  
129 tank used for washing hands at hospital C (sample BH52); **F)** tap water in a distant village in Benin (sample  
130 BSE93); and **G)** well water 100 m away from hospital A in Benin (sample BH11). **H)** In Burkina Faso, the HWW  
131 septic tank is shown for hospital F (sample BFH1). Photos by K. Haukka, A. Sarekoski and A. Butcher.



132

133 **Figure S3. PCoA plots describing the ordinations based on A) resistomes, B) taxonomical compositions,**  
 134 **and C) mobilomes present in HWWs from Benin, Burkina Faso, and Finland.** Countries are coded as colors,  
 135 and the specific hospitals are marked with a corresponding capitalized letter. Confidence ellipses represent 90 %  
 136 confidence levels.

137

138 **Table S1. Significance levels for ordination plots.** The significance was measured using equal size sample  
 139 groups (n=8/country) of randomly selected samples. The triplicated tests were mainly in line with each other,  
 140 suggesting that sample selection did not significantly affect the results. The significance levels are denoted as  
 141 follows: '\*\*\*'  $\leq 0.001$ ; '\*\*' = 0.001-0.01; '\*' = 0.01-0.05; '.' = 0.05-0.1; ' ' = 0.1-1.

ARGs, Horn									
	sample set #1			sample set #2			sample set #3		
Benin-Finland	0.29492	0.001	***	0.30982	0.002	**	0.33668	0.001	***
Benin-Burkina Faso	0.09321	0.121		0.17575	0.002	**	0.16236	0.007	**
Burkina Faso-Finland	0.24544	0.001	***	0.25117	0.001	***	0.25585	0.001	***

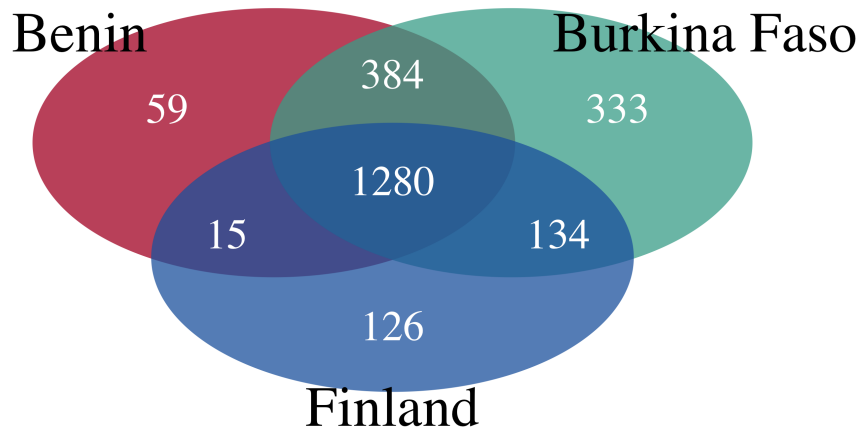
  

Taxonomical composition, Horn									
	sample set #1			sample set #2			sample set #3		
Benin-Finland	0.17282	0.002	**	0.20603	0.001	***	0.18496	0.002	**
Benin-Burkina Faso	0.08476	0.159		0.14289	0.007	**	0.08541	0.175	
Burkina Faso-Finland	0.16481	0.001	***	0.14788	0.006	**	0.16983	0.002	**

MGEs Horn									
	sample set #1			sample set #2			sample set #3		
Benin-Finland	0.31312	0.002	**	0.34187	0.001	***	0.34196	0.001	***
Benin-Burkina Faso	0.06114	0.497		0.10201	0.098	.	0.0994	0.094	.
Burkina Faso-Finland	0.20934	0.001	***	0.22952	0.001	***	0.20678	0.002	**

142  
143



144  
145  
146  
147

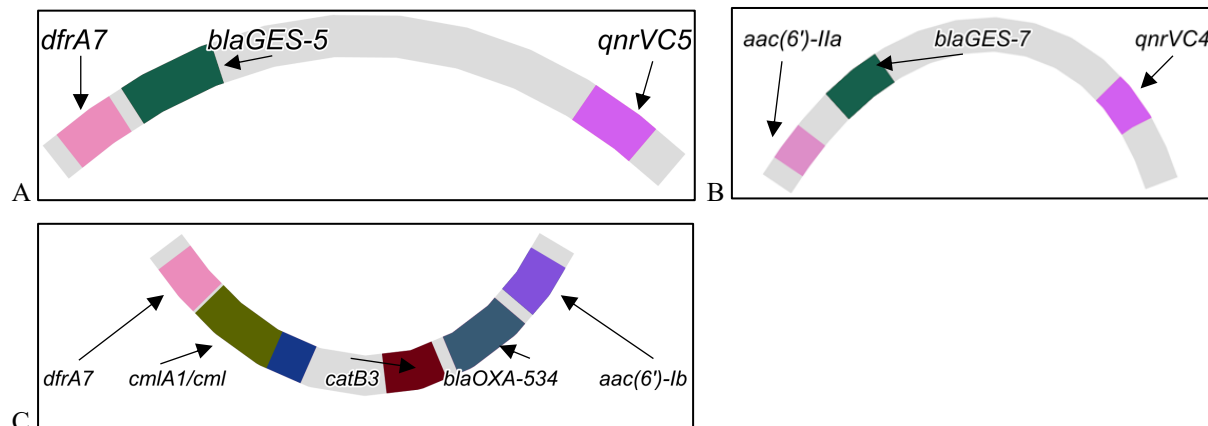
**Figure S4. Unique and shared ARGs between hospital wastewaters in Benin, Burkina Faso, and Finland.**  
A total of 2,331 ARGs were detected among 67 hospital wastewater samples in the three countries. The number of unique ARGs was substantially higher (333) in Burkina Faso compared to Benin (59) or Finland (126)



148

149 **Figure S5. Relative abundance of clinically relevant species.** The heatmap shows the relative abundances of  
 150 species listed in the ESKAPEEc group (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*,  
 151 *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter* spp., and *Escherichia coli*) defined by de  
 152 Angelis and colleagues<sup>25</sup>. In addition, *Acinetobacter* species with known intrinsic resistance to carbapenems  
 153 according to the Beta-lactamase Database<sup>26</sup> are shown. Of these, species with a total relative abundance higher  
 154 than 0.05 in the samples are shown.

155



156

157

158 **Figure S6. Examples of assembly graphs showing multidrug-resistance fragments putatively encoded by**  
 159 **class 1 integron gene cassettes in samples A) BH02, B) BH48, and C) BFH19.**

160

161



162



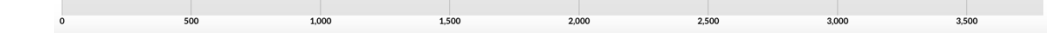
163



164



165



166



167

168

169

170

171

172

173

174 **Figure S7. Comparative analysis of contigs that harbor plasmid-mediated *mcr-5.1* colistin resistance gene**  
 175 **in four different metagenomic samples.** Visualizations of the contigs were created using anvio (v7)  
 176 metagenomics workflow for single profiles<sup>13</sup>. Putative encoding genes are indicated by blue arrows. The gene  
 177 names were determined based on BLAST search results. The recruited reads are shown as a dark grey histogram  
 178 with an overlay line (pink) denoting the GC content. Single nucleotide variation is shown on top of the panel with  
 179 colourful vertical lines. Inverse repeat sequences of Tn3-like transposase sequences (38 bp) are shown with dark  
 180 pink (IRL: 5' GGGGTCGTCTCAGAAAACGGAAAAAATCGTACGCTAAG 3'<sup>27</sup>) and green triangles (IRR: 5'  
 181 CTTAGCGTACGATTTTTTCCGTTTTCTAGACGACCCC 3'<sup>27</sup>). The genetic environments are shown for  
 182 samples **A) BH02 B) BH48, C) BFH26 D) BFH41, and E) FH9.**

183

184 *References*

- 185 1. Andrews, S. FastQC A quality control tool for high throughput sequence data  
 186 <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. (2010).
- 187 2. Ewels, P., Ns Magnusson, M., Lundin, S. & Allier, M. K. Data and text mining MultiQC: summarize  
 188 analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048.
- 189 3. Marcel, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.  
 190 *EMBnet.journal* **17**, 10–12 (2011).
- 191 4. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359  
 192 (2012).
- 193 5. Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob.*  
 194 *Chemother.* **75**, 3491–3500 (2020).
- 195 6. Pärnänen, K. *et al.* Maternal gut and breast milk microbiota affect infant gut antibiotic resistome and  
 196 mobile genetic elements. *Nat. Commun.* **9**, 389 (2018).
- 197 7. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial  
 198 communities with bioBakery 3. *Elife* **4**, 1–42 (2021).
- 199 8. Bengtsson-Palme, J. *et al.* metaxa2: Improved identification and taxonomic classification of small and  
 200 large subunit rRNA in metagenomic data. *Mol. Ecol. Resour.* **15**, 1403–1414 (2015).
- 201 9. Eddy, S. R. Accelerated Profile HMM Searches. *Cit. Eddy SR* **7**, 1002195 (2011).
- 202 10. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, 427–432 (2018).
- 203 11. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution  
 204 for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–  
 205 1676 (2015).
- 206 12. Altschup, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search  
 207 Tool. *J. Mol. Biol* **215**, 403–410 (1990).
- 208 13. Murat Eren, A. *et al.* Community-led, integrated, reproducible multi-omics with anvio. *Nat. Microbiol.*  
 209 **6**, 3–6 (2021).
- 210 14. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC*  
 211 *Bioinformatics* **11**, (2010).
- 212 15. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. Sequence analysis CD-HIT: accelerated for clustering the  
 213 next-generation sequencing data. **28**, 3150–3152 (2012).

- 214 16. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo  
215 genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
- 216 17. Wickham, H. ggplot2: Elegant graphics for data analysis. *Springer-Verlag New York*  
217 <https://ggplot2.tidyverse.org> (2016).
- 218 18. Kolde, R. pheatmap: Pretty Heatmaps. R package version 1.0.12. [https://cran.r-](https://cran.r-project.org/package=pheatmap)  
219 [project.org/package=pheatmap](https://cran.r-project.org/package=pheatmap) (2019).
- 220 19. Pedersen, T. L. patchwork: The Composer of Plots. R package version 1.1.1. [https://cran.r-](https://cran.r-project.org/package=patchwork)  
221 [project.org/package=patchwork](https://cran.r-project.org/package=patchwork) (2020).
- 222 20. South, A. rnaturalearth: World Map Data from Natural Earth. R package version 0.1.0. [https://cran.r-](https://cran.r-project.org/package=rnaturalearth)  
223 [project.org/package=rnaturalearth](https://cran.r-project.org/package=rnaturalearth).
- 224 21. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and  
225 Graphics of Microbiome Census Data. *PLoS One* **8**, (2013).
- 226 22. Oksanen, J. *et al.* vegan: Community Ecology Package. <https://cran.r-project.org/package=vegan>  
227 (2020).
- 228 23. Chen, H. VennDiagram: Generate High-Resolution Venn and Euler Plots. R package version 1.6.20.  
229 <https://cran.r-project.org/package=VennDiagram> (2018).
- 230 24. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq  
231 data with DESeq2. *Genome Biol.* **15**, 550 (2014).
- 232 25. De Angelis, G. *et al.* Incidence and antimicrobial resistance trends in bloodstream infections caused by  
233 ESKAPE and *Escherichia coli* at a large teaching hospital in Rome, a 9-year analysis (2007-2015). *Eur.*  
234 *J. Clin. Microbiol. Infect. Dis.* **37**, 1627–1636 (2007).
- 235 26. Naas, T. *et al.* Beta-lactamase database (BLDB)-structure and function. *J. Enzyme Inhib. Med. Chem.*  
236 **32**, 917–919 (2017).
- 237 27. Borowiak, M. *et al.* Identification of a novel transposon-associated phosphoethanolamine transferase  
238 gene, *mcr-5*, conferring colistin resistance in d-tartrate fermenting *Salmonella enterica* subsp. *enterica*  
239 serovar Paratyphi B. *J. Antimicrob. Chemother.* **72**, 3317–3324 (2017).
- 240