The extent of carbapenemase encoding genes in public genome sequences

5 Ingmar Janse[1*], Rick Beeloo[1*], Arno Swart[1], Michael Visser[2], Leo Schouls[1], Engeline van

6 Duijkeren[1], Mark W. J. van Passel[1,3]

8 [1] National Institute for Public Health and the Environment, Center for Infectious Disease Control,

9 Bilthoven, the Netherlands

10 [2] Netherlands Food and Consumer Product Safety Authority (NVWA), Sequencing and

11 Bioinformatics, Utrecht, the Netherlands

12 [3]Ministry of Health, Welfare and Sport, The Hague, the Netherlands

14 *shared first authorship

16 Corresponding Author:

17 Ingmar Janse, National Institute for Public Health and the Environment, Center for Infectious

18 Disease Control, Bilthoven, the Netherlands

19 Email address: ingmar.janse@rivm.nl

20

Abstract

Genome sequences provide information on the genetic elements present in an organism, and currently there are databases containing hundreds of thousands of bacterial genome sequences. These repositories allow for mining patterns concerning antibiotic resistance gene occurrence in (pathogenic) bacteria in e.g. natural environments or animal reservoirs, and link these to relevant metadata such as bacterial host species, year and country of isolation, and co-occurrence with other resistance genes. In addition, the advances in the prediction of mobile genetic elements, and discerning chromosomal from plasmid DNA, broadens our view on the mechanism mediating dissemination.

In this study we utilize the vast amount of data in the public database PATRIC to investigate the dissemination of carbapenemase-encoding genes (CEGs), the emergence and spread of which is considered a grave public health concern.

Based on publicly available genome sequences from PATRIC and manually curated CEG sequences from the beta lactam database, we found 7,964 bacterial genomes, belonging to at least 70 distinct species, that carry in total 9,892 CEGs, amongst which bla_{NDM}, bla_{OXA}, bla_{VIM}, bla_{IMP} and bla_{KPC}. We were able to distinguish between chromosomally located resistance genes (4,137; 42%) and plasmid-located resistance genes (5,753; 58%). We found that a large proportion of the identified CEGs were identical, i.e. displayed 100% nucleotide similarity in multiple bacterial species (8,361 out of 9,892 genes; 85%). For example, the New Delhi metallo-beta-lactamase NDM-1 was found in 42 distinct bacterial species, and present in seven different reservoirs. Our data show the extent of carbapenem-resistance far beyond the canonical species *Acetinobacter baumannii, Klebsiella pneumoniae* or *Pseudomonas aeruginosa*.

These types of data complement previous systematic reviews, in which carbapenem-resistant Enterobacteriaceae were found in wildlife, livestock and companion animals. Considering the widespread distribution of CEGs, we see a need for comprehensive surveillance and transmission studies covering more host species and environments, akin to previous extensive surveys that focused on extended spectrum beta-lactamases. This may help to fully appreciate the spread of CEGs and improve the understanding of mechanisms underlying transmission, which could lead to interventions minimizing transmission to humans.

Introduction

Twenty-five years since the sequencing revolution started we see the availability of hundreds of thousands of bacterial genomes [1], enabling a vast number of comparative analyses leading to e.g. virulence factor prediction [2] or metabolic pathway modelling [3]. The extensive genome resource PATRIC [1] includes detailed information on resistance genes [4], and amongst the many different uses for these accumulated data, they also aim at predicting antimicrobial resistance (AMR) phenotypes in individual genomes. However, the accumulation of all genetic data with their associated metadata on bacterial host species, isolation source, year of isolation, and co-occurrence with other clinical resistance genes, also allows for a global view into the dissemination of resistance genes, for example of carbapenemase-encoding genes in high priority pathogens [5].

Recently, Köck et al. [6] performed a systematic literature review on the occurrence of carbapenem-resistant Enterobacteriaceae (CRE). The authors concluded that the prevalence of CRE in wildlife, livestock and companion animals and directly exposed humans poses public health risks, which was further corroborated by a study on the worsening epidemiological situation in Europe of carbapenemase-producing Enterobacteriaceae [7]. The literature review by Köck et al. was limited to Enterobacteriaceae, and environmental prevalence of CRE had not been included, even though genes encoding most acquired carbapenemases are believed to have transferred from environmental bacteria into species with clinical relevance [8]. We propose that evaluating carbapenemase-encoding genes and their associated metadata from genomic databases may add to the broader picture of the distribution of these genes throughout relevant (non-human) reservoirs, and may highlight transmission between bacterial species via mobile genetic elements. These databases can show whether other bacterial species can represent CEG reservoirs, even when the original researchers of these genomes may not have annotated the carbapenem-resistance potential of these species. We aim to tabulate carbapenemase-encoding genes from an extensive genome database, in order to appraise their distribution patterns.

Materials & Methods

78
79

80 All available 175,882 bacterial genomes (August 2018) and their corresponding metadata were
81 downloaded from PATRIC [1]. This metadata included information on the isolated bacteria such
82 as species, taxonomic lineage, isolation source (host organisms, e.g. human, or environment, e.g.
83 forest or wastewater).
84 In addition, the nucleotide (.ffn) gene and protein sequences (.faa) were retrieved from the FTP
85 server (ftp.patricbrc.org). The taxonomic lineages for each genome were re-annotated at the
86 species and genus level by deriving these from the genome name. Species names were identified
87 by removing strain specifications after the 'sp.' suffix and omitting the 'subsp.' Suffix, in addition
88 to manual curation. Isolation information for each genome is recorded in free-text fields that
89 complicates unambiguous analysis.
90 In order to bin isolation source metadata from all genomes correctly, host names were queried
91 against the NCBI taxonomy database to retrieve taxonomic lineages and allow assignment to
92 groups such as cattle, chicken, and human. Source data from environmental or food isolates were
93 standardized by querying these against manually curated keywords (Table S1) to assign each
94 genome to specified reservoirs (Fig. S2). In case source metadata was absent, the genome was
95 designated "miscellaneous".

96

97 To create a reliable antibiotic resistance gene (ARG) set, protein sequences of all PATRIC ARGs
98 identified upon comparison with CARD (indicated by the "source" column in PATRIC annotation)
99 were subjected to the Resistance Gene Identifier (RGI [9] and only those detected by the protein
100 homolog model with a perfect or strict cut-off were kept. To identify carbapenemase-encoding
101 genes (CEGs) in this dataset, the protein sequences of 566 manually curated CEGs (Data S3) were
102 obtained from the Beta-lactamase database (BLDB) [10] and detected in the ARG set using
103 TBLASTN (100% identity and coverage). The CEGs were clustered on 100% sequence identity
104 and coverage using Python, and their genomic locations (chromosome or plasmid) were predicted
105 using MOB-RECON [11].

106
107
108
109
110
111
112
113
114

Results

115
116

117 ***Carbapenemase encoding genes in bacterial genomes***
118 From the collection of nearly 2 million antibiotic resistance genes in over 170,000 bacterial
119 genomes, we identify 9,892 carbapenemase encoding genes (CEGs) in 7,964 bacterial genomes.
120 The set of 9,892 CEGs consisted of 203 unique protein sequences and 257 unique nucleotide
121 sequences, respectively. The former reflect the amino-acid based beta-lactam classifications
122 system used by the BLDB [10] that allowed accurate assignment of each CEG to its protein name.
123 Genes encoding KPC-2, OXA-23, OXA-66, and KPC-3 were detected over 1000 times in our
124 dataset. The 7,964 genomes belonged to at least 77 bacterial species (18 lacked taxonomic
125 assignment beyond the genus level, Table S4), from 33 genera, 18 families and 5 phyla.

126 ***Multiple occurrences of CEGs within a single genome***
127 We found that while the majority (6,191) of genomes only encoded a single CEG, 1,773 genomes
128 carried multiple CEGs. Of the genomes carrying multiple CEGs, 189 harboured CEGs encoding
129 the same protein, with OXA-23 encoding genes being most frequently present in multiple copies
130 (41 genomes), followed by KPC-3 encoding genes (n=35), KPC-2 encoding genes (n=30), NDM-1
131 encoding genes (n=12), and OXA-72 encoding genes (n=9). Interestingly, we found 38 genomes
132 containing three or more genes encoding the same protein, and three genomes with five copies for
133 the same protein, KPC-2 for *Klebsiella michiganensis*, and OXA-23 and OXA-72 for
134 *Acinetobacter baumannii*. Moreover, we also found multiple gene copies encoding for the same
135 protein in less common pathogens, such as three copies encoding KPC-3 in *Serratia marcescens*
136 and *Raoultella planticola*.

137 Most of the genomes with multiple CEGs carried genes encoding different proteins (n = 1,666),
138 which were of the same carbapenemase type for the majority of these genomes (n= 1,533).
139 Combinations of OXA (oxacillinase) type CEGs were most common, especially bla_{OXA-23} and
140 bla_{OXA-66} (803 genomes) and bla_{OXA-23} and bla_{OXA-82} (290 genomes). However, 133 genomes
141 contained CEGs encoding carbapenemases of two different types. The most frequent combination
142 of genes encoding distinct carbapenemase types, i.e. bla_{NDM-1} and $bla_{OXA-232}$ were found in 25
143 *Klebsiella pneumoniae* genomes and 1 *Escherichia coli* genome. Other common combinations
144 were bla_{NDM-1} and bla_{OXA-94} in 11 *Acinetobacter baumannii* genomes, bla_{KPC-2} and bla_{NDM-1} in 5
145 *Enterobacter cloacae*, 2 *Klebsiella pneumoniae*, 2 *Klebsiella michiganensis*, and 1 *Citrobacter*
146 *freundii*, and lastly bla_{NDM-1} and bla_{OXA-23} in 10 *Acinetobacter baumannii* genomes, 1
147 *Acinetobacter radioresistens* genome and 1 *Acinetobacter* sp. genome.

148 ***Distribution of CEGs between species and reservoirs***
149 Most CEGs were found in the genera *Acinetobacter* (3074 genomes : 4783 genes), *Klebsiella*
150 (3267 : 3404), *Escherichia* (542 : 559), *Enterobacter* (485 : 502) and *Pseudomonas* (268 : 281).
151 Overall, CEGs were found most frequently present in the species *Acinetobacter baumannii* (4,502
152 CEGs), *Klebsiella pneumoniae* (3,285 CEGs), *Escherichia coli* (557 CEGs), *Enterobacter cloacae*
153 (313), *Pseudomonas aeruginosa* (245 CEGs), and also, albeit to less extent, in species less
154 commonly included in carbapenemase surveillance such as *Serratia marcescens* (66 CEGs),

155 *Citrobacter freundii* (59 CEGs), *Klebsiella oxytoca* (36), *Klebsiella quasipneumoniae* (27),
156 *Klebsiella michiganensis* (24 CEGs), and *Proteus mirabilis* (24 CEGs), and *Acinetobacter*
157 *radioresistens* (20 CEGs). The important CEG bla_{IMP-1} was most often detected in *Enterobacter*
158 *hormaechei* and *Klebsiella pneumoniae,* but also in *Acinetobacter baumannii, A. junii, A. pittii, A.*
159 *nosocomialis, Enterobacter cloacae, E. kobei, E. asburiae, Escherichia coli, Pseudomonas*
160 *aeruginosa* and *Serratia marcescens.* bla_{OXA48} was predominantly found in *K. pneumoniae*, but
161 also in *Escherichia coli, Enterobacter aerogenes, E. cloacae, E. hormaechei, E. kobei, Citrobacter*
162 *freundii, C. koseri, Shewanella* sp. (on chromosome), *Kluyvera ascorbata, Proteus mirabilis* and
163 *Raoultella ornithinolytica.* $bla_{NDM-1,}$ bla_{KPC-2} and bla_{KPC-3} were most often identified in *K.*
164 *pneumoniae*, followed by *E. coli* and *E. cloacae* (Table S5, S6 and S7).

165 Nucleotide sequence comparison of the 9,892 CEGs revealed that 1,591 gene sequences were
166 present in only a single bacterial species, whereas the remaining 8,301 gene sequences were found
167 identically in multiple species. For example, the well-known CEGs bla_{NDM-1}, bla_{KPC-2}, bla_{IMP-1} and
168 bla_{OXA-48} in at least 38, 27, 12 and 13 species, respectively.

169 Based on the isolation information for each genome we distinguished 11 different reservoirs of
170 CEGs, of which the human isolates represented by far the largest reservoir in this study (8,099
171 genes, 82%). CEGs such as bla_{KPC-2}, bla_{OXA-23} and bla_{OXA-66} were dominant when considering all
172 environments. However, when the human and miscellaneous (which will likely contain many
173 human isolates) genomes were excluded from the dataset, dominant CEGs shifted, leading to high
174 frequencies of bla_{NDM-5}, bla_{NDM-9} and bla_{NDM-1}, mostly identified in bacteria isolated from chickens
175 (109 genomes). The distribution of CEGs over the different species and environments is presented
176 in Figure 1.

177 ***Genomic location of CEGs***
178 We evaluated the genomic location of CEGs based on their genomic context and found that 4,137
179 (42%) were predicted to be chromosomally located while 5,753 (58%) were predicted to be
180 plasmid-located. bla_{IMP1} was most often located on IncN or IncW plasmids, bla_{KPC-2} on IncFII or
181 IncU plasmids, bla_{KPC-3} on IncFIA or IncFII plasmids, bla_{NDM-1} on IncA/C plasmids, bla_{NDM-5} on
182 IncX3 plasmids and bla_{OXA-48} on Inc L/M plasmids.
183
184
185
186
187

Discussion

The identification of CEGs in a large number of genomes of a diversity of bacterial species showed that the distribution of these high-priority resistance genes extends far beyond the well-known clinically relevant carbapenem-resistant *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacteriaceae,* which are known to be emerging causes for hospital acquired infections (https://www.who.int/infection-prevention/publications/guidelines-cre/en/).

A recent cross-border collaborative effort on whole genome sequencing of certain carbapenemase-positive *K. pneumoniae* identified muli-country transmission clusters [12]. This analysis relied on 143 genome sequences, and underlined the benefits of data sharing. The integration of public data sources, such as the PATRIC genome repository with a backdrop of over 3,200 carbapenemase containing *K. pneumoniae* genomes, in addition to a substantive potential reservoir of mobile carbapenemase-encoding genes, may provide further context and epidemiological links for both strains as well as mobile elements outside the regularly surveilled species.

A considerable number of these genomes carried two or more different CEG genes (Table 1). A consequence of co-expression of different types of carbapenemase genes, for example $bla_{\text{OXA-48}}$ and $bla_{\text{NDM-1}}$, together with other resistance genes that might be present on the plasmids, is a reduction of treatment options [13]. In this study, only carbapenemase-encoding genes were included in the analyses, but their co-occurrence with antibiotic resistance genes from completely different classes, such as aminoglycosides, tetracyclines, et cetera, may be of interest as well to evaluate the emergence of multidrug resistant pathogens.

We found that identical CEGs are frequently shared between bacterial species and -as a consequence- can be encountered in a large number of environments (Figure 1). Horizontal gene transfer is an important evolutionary force for adaptation, and mobile genetic elements are well-known vectors for gene exchange, especially for antibiotic resistance genes [14]. The high percentage (58%) of predicted plasmid-located CEGs in the genomes shows that a large proportion of CEGs have an enhanced capacity for horizontal gene transfer. Previous genomic epidemiology studies may not have fully appreciated the widespread character of CEG dissemination [15], and possible co-mobility with other resistance determinants.

The database approach identified human isolates as the largest reservoir of CEGs. This however does not necessarily mean that the human reservoir is the dominant reservoir for carbapenem-resistance genes, but rather may reflect the frequent sampling of humans. In addition, it is likely that many more CEG-reservoir combinations exist as this analysis is limited to the isolates present in the database. Most non-human reservoirs also contained more than one type of CEG (Table 2).

228 **Conclusions**

229

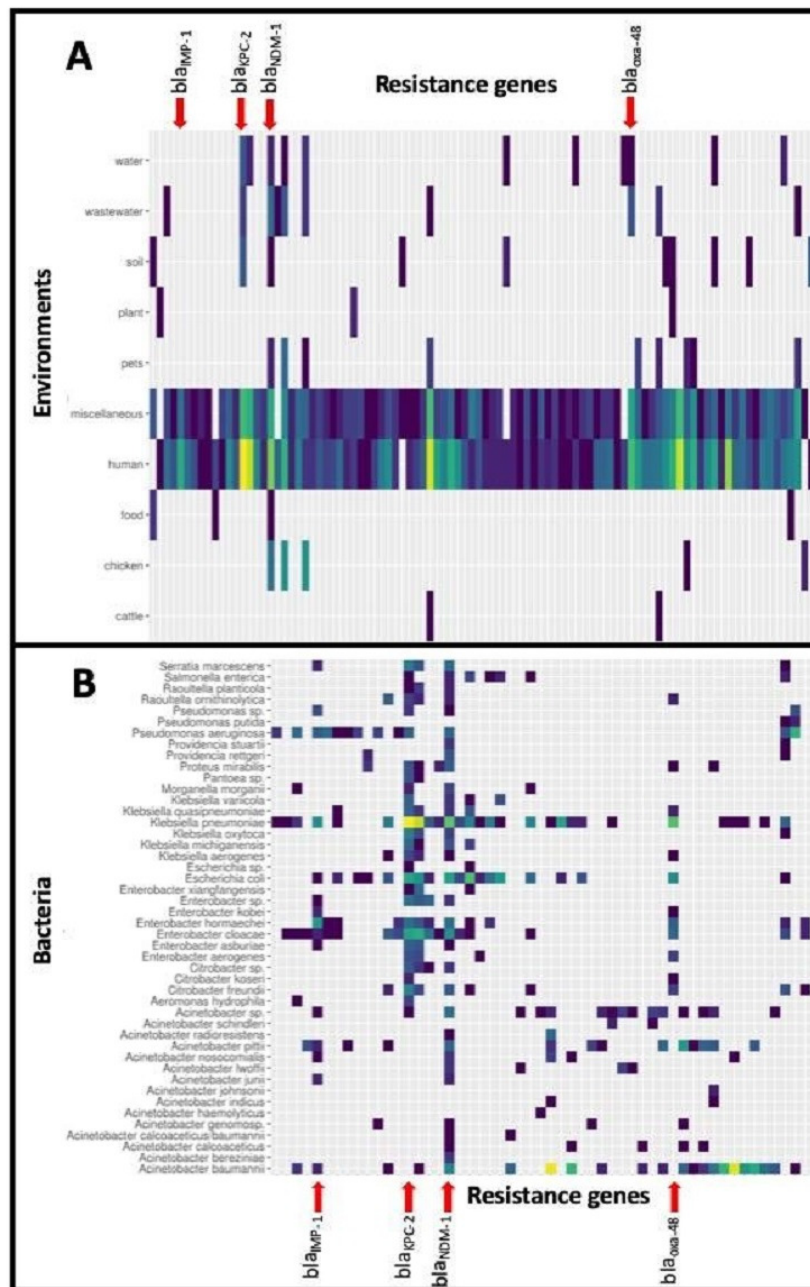
230 Large genome repositories are treasure troves for comparative analyses in different fields of study,
231 even though these databases do not represent surveillance data and therefore contain bacterial
232 isolates at other frequencies compared to their isolation sources, nor are they free of annotation
233 errors. Still, owing to their vast sizes, genome sequence collections allow for an alternative view
234 on the distribution of resistance genes and exchange potential between bacteria and reservoirs.
235 Carbapenemase-encoding genes are distributed over a large number of bacterial species, with some
236 genes, such as bla_{NDM-1}, present in at least 42 different species, isolated from at least seven distinct
237 non-human reservoirs. These numbers show the breadth of reservoirs and the vast potential for
238 resistance gene exchange and transmission. Data sharing and the use of open repositories for
239 surveillance purposes would therefore benefit evaluations of health risks for resistant infections.
240

References

1. Wattam, A.R., et al., *Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center.* Nucleic Acids Res, 2017. **45**(D1): p. D535-D542.
2. Mao, C., et al., *Curation, integration and visualization of bacterial virulence factors in PATRIC.* Bioinformatics, 2015. **31**(2): p. 252-8.
3. Karp, P.D., et al., *Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.* Brief Bioinform, 2010. **11**(1): p. 40-79.
4. Antonopoulos, D.A., et al., *PATRIC as a unique resource for studying antimicrobial resistance.* Brief Bioinform, 2019. **20**(4): p. 1094-1102.
5. Rello, J., et al., *A global priority list of the TOp TEn resistant Microorganisms (TOTEM) study at intensive care: a prioritization exercise based on multi-criteria decision analysis.* Eur J Clin Microbiol Infect Dis, 2019. **38**(2): p. 319-323.
6. Kock, R., et al., *Carbapenem-resistant Enterobacteriaceae in wildlife, food-producing, and companion animals: a systematic review.* Clin Microbiol Infect, 2018. **24**(12): p. 1241-1250.
7. Brolund, A., et al., *Worsening epidemiological situation of carbapenemase-producing Enterobacteriaceae in Europe, assessment by national experts from 37 countries, July 2018.* Euro Surveill, 2019. **24**(9).
8. Woodford, N., et al., *Carbapenemase-producing Enterobacteriaceae and non-Enterobacteriaceae from animals and the environment: an emerging public health risk of our own making?* J Antimicrob Chemother, 2014. **69**(2): p. 287-91.
9. Jia, B., et al., *CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database.* Nucleic Acids Res, 2017. **45**(D1): p. D566-D573.
10. Naas, T., et al., *Beta-lactamase database (BLDB) - structure and function.* J Enzyme Inhib Med Chem, 2017. **32**(1): p. 917-919.
11. Robertson, J. and J.H.E. Nash, *MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies.* Microb Genom, 2018. **4**(8).
12. Ludden, C., et al., *Cross-border spread of bla NDM-1- and bla OXA-48-positive Klebsiella pneumoniae: a European collaborative analysis of whole genome sequencing and epidemiological data, 2014 to 2019.* Euro Surveill, 2020. **25**(20).
13. Gona, F., et al., *Emergence of two novel sequence types (3366 and 3367) NDM-1- and OXA-48-co-producing K. pneumoniae in Italy.* Eur J Clin Microbiol Infect Dis, 2019. **38**(9): p. 1687-1691.
14. Ding, B., et al., *In vivo Acquisition of Carbapenemase Gene blaKPC-2 in Multiple Species of Enterobacteriaceae through Horizontal Transfer of Insertion Sequence or Plasmid.* Front Microbiol, 2016. **7**: p. 1651.
15. Wilson, H. and M.E. Torok, *Extended-spectrum beta-lactamase-producing and carbapenemase-producing Enterobacteriaceae.* Microb Genom, 2018. **4**(7).

Figure 1

Distribution of the 9,892 carbapenemase-encoding genes in bacterial species and non-human reservoirs.

Horizontal lines represent reservoirs (A) or bacterial species (B) that collectively harbor different CEGs. Vertical lines represent CEGs that occur in many different bacterial species or reservoirs: four genes have been highlighted. The figure was simplified by removing the rare CEGs, and the full data is available in Supplementary Table S5, S6 and S7.

The distribution of the 9,892 carbapenemase-encoding genes (CEG) in 7,884 bacterial genomes. The majority of genes (6,191) is present in a single copy in its genome, but 3,701 genes reside in a genome with additional carbapenemase-encoding genes

		Number of different CEGs		
		1	**2**	**3**
Number of CEGs per genome	**1**	**6,191**	0	0
	2	81	**1,567**	0
	3	21	52	**32**
	4	4	8	0
	5	1	4	1
	6	0	2	0

Diversity of CEGs in the 11 reservoirs from the genome database

Reservoir	# Distinct CEGs	CEGs	# Species	Species*
Pig	1	NDM-1	1	*A. baumannii*
Cattle	3	OXA-146,OXA-23,OXA-58	1	*A. indicus*
Plant	3	CGB-1,OXA-120,OXA-65	2	*A. baumannii, C. indologenes,*
Chicken	5	NDM-1,NDM-5,NDM-9,OXA-68,VIM-48	5	*A. baumannii, E. coli, K. pneumoniae, S. enterica, P. putida*
Food	5	BcII-1,IMP-27,NDM-1,OXA-497,VIM-1	6	*A. baumanni, B. cereus, B. thuringensis, E. coli, V. parahaemolyticus, S. enterica*
Pets	9	NDM-1,NDM-5,NDM-9,OXA-23,OXA-500,OXA-58,OXA-68,OXA-69,VIM-2	6	*A. baumannii, A. gandensis, A. pittii, A. radioresistens, E. coli, P. aeruginosa*
Soil	10	BcII-1,KPC-2,NDM-1,OXA-213,OXA-273,OXA-64,OXA-65,OXA-72,OXA-91,VIM-5	10	*A. baumannii, A. calcoaceticus, Acinetobacter sp., B. cereus, B. subtilis, C. braakii, C. freundii, Erythrobacter sp., P. guariconensis, P. plecoglossicida*
Wastewater	11	GES-5,KPC-2,NDM-1,NDM-4,NDM-5,NDM-9,OXA-23,OXA-48,OXA-58,SHV-38,VIM-2	11	*A. johnsonii, A. junii, Acinetobacter sp., A. hydrophila, C. freundii, E. cloacae, E. kobei, E. coli, Escherichia sp., K. pneumoniae, Pseudomonas sp.*
Water	13	KPC-2,KPC-3,NDM-1,NDM-5,NDM-9,OXA-269,OXA-273,OXA-280,OXA-360,OXA-444,OXA-48,OXA-72,SPM-1	14	*A. calcoaceticus, A. johnsonii, A. pittii, A. schindleri, E. cloacae, E. coli, G. pentaromativorans, K. pneumoniae, K. quasipneumoniae, K. variicola, P. aeruginosa, Pseudomonas sp., Ralstonia sp., V. cholerae*
Miscellaneous	109		48	***
Human	175		73	***

3 *) for the full names, please refer to Supplementary Table S3

4 **) the list of genes is in the Supplementary Table S4

5 ***) the list of species is in the Supplementary Table S5