The resistome of important human pathogens

Christian Munck, Mostafa Ellabaan, Michael Schantz Klausen, Morten O.A. Sommer*

The Novo Nordisk Foundation Center for Biosustainability
Technical University of Denmark
Kogle Alle 6, 2970 Hørsholm, Denmark
*msom@bio.dtu.dk

Abstract

Genes capable of conferring resistance to clinically used antibiotics have been found in many different natural environments. However, a concise overview of the resistance genes found in common human bacterial pathogens is lacking, which complicates risk ranking of environmental reservoirs. Here, we present an analysis of potential antibiotic resistance genes in the 17 most common bacterial pathogens isolated from humans. We analyzed more than 20,000 bacterial genomes and defined a clinical resistome as the set of resistance genes found across these genomes. Using this database, we uncovered the co-occurrence frequencies of the resistance gene clusters within each species enabling identification of co-dissemination and co-selection patterns. The resistance genes identified in this study represent the subset of the environmental resistome that is clinically relevant and the dataset and approach provides a baseline for further investigations into the abundance of clinically relevant resistance genes across different environments. To facilitate an easy overview the data is presented at the species level at www.resistome.biosustain.dtu.dk.

56 **Introduction**

57 Over the past decade substantial efforts have been devoted to characterize the
58 antibiotic resistome in different environments[1,2]. Genes conferring resistance to
59 most antibiotics have been found in most environments[1], including pristine
60 environments such as permafrost sediments and isolated cave environments[3,4].
61 In addition, several clinically relevant resistance genes have recently emerged in
62 human pathogens highlighting the flow of resistance genes from natural
63 environments to human pathogens[5]. Yet, an increasing number of studies
64 suggest that substantial barriers to gene transfer across environmental niches
65 exist, implying that genes capable of conferring antibiotic resistance are not
66 easily transferred to human pathogens, especially if selection is absent [6-8]. In
67 order to assess the overlap between an environmental resistome and clinically
68 relevant resistance genes a general and comprehensive overview of resistance
69 gene prevalence across human pathogenic species is required. Currently, such
70 overview is only available for subsets of resistance mechanisms or species and
71 have relied on extensive literature mining [9-13] Yet, with the advance of affordable
72 whole-genome sequencing, multiple large-scale sequencing studies of key human
73 pathogens have been conducted[14,15]. The data generated in such studies give a
74 detailed insight into the evolution and dissemination of pathogenic strains and
75 enable researchers to develop species-specific genome-based predictions of the
76 resistance genes underlying the resistance phenotypes[16-21]. Still, an unbiased
77 analysis of the resistome of pathogenic human isolates is lacking.

78

79 In order to get a more complete overview of the antibiotic resistance gene
80 distribution in important human pathogens, we have analyzed all available
81 genomes obtained from human isolates of species commonly known to cause
82 infections. Our study includes genomes available for all the ESKAPE pathogens[22]
83 (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*,
84 *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter spp*.) as well
85 as *Escherichia coli*, *Mycobacterium tuberculosis*, *Neisseria gonorrhoeae*, *Neisseria*
86 *meningitidis*, *Streptococus pneumoniae*, *Streptococcus pyogenes*, *Enterococcus*
87 *faecalis*, *Salmonella enterica*, *Shigella flexneri* and *Campylobacter jejuni*. These

88 species represent the most common human bacterial pathogens, and antibiotic
89 resistance within many of them is increasing [23].
90
91 Based on our genomic analysis we defined a set of clinically relevant resistance
92 genes found in these species. Our analysis provides a comprehensive overview of
93 resistance genes across different pathogens and we provide an easy accessible
94 presentation of the data at www.resistome.biosustain.dtu.dk. This subset of
95 resistance genes will make it easier to identify clinically relevant resistance
96 genes in environmental sequence data and uncover connections between
97 different environmental resistance gene reservoirs and the clinic.
98
99
100 **Results**
101
102 We downloaded all 31,073 genomes available for the ESKAPE pathogens plus *E.*
103 *coli*, *M. tuberculosis*, *N. gonorrhoeae*, *N. meningitides*, *S. pneumoniae*, *S. pyrogenes*,
104 *E. faecalis*, *S. enterica*, *S. flexneri* and *C. jejuni* (table 1). From the host descriptor
105 in the genbank files we identified 20,757 (67%) genomes that derived from
106 human isolates. Using BLAST we mined the genomes for resistance genes with
107 the highly curated Resfinder database[24] as the query sequences (Materials and
108 methods).
109
110 With this approach we identified 122,463 putative resistance genes across the
111 genomes. Many of these genes represented different variants of the same gene
112 such as the different *tem* or *oxa* betalactamase genes. In order to simplify the
113 results we clustered the resistance genes at 80 % identity (complete un-
114 clustered dataset with gene variant resolution is available on
115 www.resistome.biosustain.dtu.dk). We named the resistance gene clusters with
116 a cluster id followed by a common name for the genes in the cluster, e.g. c205-
117 TEM is cluster 205, which contains the *tem* beta-lactamase genes.
118
119 In total, 197 unique resistance gene clusters were identified. Of these, some were
120 identified as being native in at least one of the species, e.g. the cluster containing

121 the *oxa-51* beta-lactamase is naturally occurring in *A. baumannii*[25]. In order to
122 exclude such native resistance genes we first identified gene clusters with high
123 species abundance (>93% of isolates of a given species) and subsequently mined
124 the literature for evidence that these gene clusters represented native genes
125 (Supplementary table 3). In total 16 gene clusters were removed even though
126 they may contribute to resistance. Importantly, they were only removed from the
127 species they were native to.
128 After removing the native resistance gene clusters 187 unique resistance gene
129 clusters remained (table 1 and supplementary table 3
130 www.resistome.biosustain.dtu.dk). These clusters represent 38 % of the 492
131 resistance gene clusters in the ResFinder database, highlighting that the majority
132 of the resistance genes clusters in the database are not found in sequenced
133 genomes of common human pathogens.
134
135
136
137
138

Species	Total genomes	Human isolates	Unique cluster	Top 5 gene clusters in human isolates
Gram negative				
Salmonella enterica	4750	1036 (22%)	62	c146-sul(12%), c133-aadA(11%), c205-TEM(9%), c232-strB(8%), c215-sul(7%)
Escherichia coli	4483	2546 (57%)	76	c205-TEM(39%), c215-sul(32%), c270-strA(29%), c232-strB(29%), c146-sul(25%)
Pseudomonas aeruginosa	1612	598 (37%)	47	c146-sul(8%), c45-aac(6')Ib-cr(6%), c424-aadB(3%), c274-OXA(2%), c133-aadA(2%)
Acinetobacter baumannii	1525	1025 (67%)	55	c270-strA(52%), c232-strB(52%), c146-sul(50%), c245-OXA(49%), c133-aadA(49%)
Klebsiella pneumoniae	1145	798 (70%)	83	c45-aac(6')Ib-cr(58%), c205-TEM(55%), c146-sul(54%), c133-aadA(49%), c232-strB(48%)
Neisseria meningitidis	724	174 (24%)	6	c65-tetB(60%), c17-tetM(1%), c151-ROB(1%), c205-TEM(1%), c232-strB(1%)
Campylobacter jejuni	646	459 (71%)	7	c301-OXA(68%), c16-tetO(29%), c283-aph(3')-III(5%), c198-aadE(4%), c42-aac(6')-aph(2'')(3%)
Enterobacter cloacae	483	122 (25%)	65	c397-fosA(89%), c91-ACT(75%), c133-aadA(40%), c270-strA(38%), c232-strB(37%)
Neisseria gonorrhoeae	329	308 (94%)	5	c19-penA(44%), c17-tetM(7%), c205-TEM(5%)

Shigella flexneri	132	120 (91%)	30	c360-catA(88%), c234-OXA(84%), c65-tetB(83%), c449-dfrA(82%), c215-sul(58%)
Enterobacter aerogenes	117	99 (85%)	37	c205-TEM(9%), c45-aac(6')Ib-cr(9%), c146-sul(7%), c133-aadA(6%), c203-aac(3)-II(6%)
Gram positive				
Streptococcus pneumoniae	7122	4457 (63%)	17	c17-tetM(55%), c41-msrD(17%), c62-mefA(17%), c310-ermB(14%), c369-cat(pC194)(8%)
Staphylococcus aureus	6854	4923 (72%)	42	c5-mecA(88%), c220-blaZ(70%), c295-spc(50%), c335-ermA(49%), c304-aadD(44%)
Enterococcus faecalis	439	261 (59%)	34	c310-ermB(58%), c17-tetM(52%), c42-aac(6')-aph(2'')(46%), c156-aadE(23%), c283-aph(3')-III(22%)
Enterococcus faecium	419	192 (46%)	39	c310-ermB(76%), c156-aadE(67%), c283-aph(3')-III(64%), c137-vanH(47%), c352-vanX(47%)
Streptococcus pyogenes	293	250 (85%)	5	c17-tetM(9%), c310-ermB(2%), c335-ermA(1%), c41-msrD(1%), c62-mefA(1%)
Mycobacterium tuberculosis	3531	3389 (96%)	0	NA

Table 1. The five most abundant resistance gene clusters in the 17 human pathogens along with general statistics on the species. The total genomes column denotes to total number of refseq genomes for each species. The human isolates column denotes the number of genomes annotated as being from a human host. The unique cluster column denotes the number of unique resistance gene clusters found for each species. Only *M. tuberculosis* has no acquired resistance genes. For complete overview visit www.resistome.biosustain.dtu.dk.

Only *M. tuberculosis* did not have any putative resistance genes, which is to be expected, as antibiotic resistance in *M. tuberculosis* is achieved through mutations in native genes as opposed to acquisition of resistance genes[26]. The remaining pathogenic species investigated had between 5-62 gene clusters (Table 1). For Gram-negative species the beta-lactamase resistance gene cluster c205-TEM and the sulfonamide resistance gene clusters c146-sul and c215-sul are the most abundant resistance gene clusters. In *E. coli* for instance, 39 % of the genomes were found to carry a beta-lactamase gene belonging to the *tem* family, while c146-sul and c215-sul was found in 25 % and 32 % of the genomes, respectively. Clusters c146-sul and c215-sul contain the *sul1* and *sul2* genes,

159 respectively. For the Gram-positive species, the tetracycline resistance cluster
160 c17-tetM and the macrolide resistance gene clusters c335-ermA and c310-ermB
161 are the most abundant resistance gene clusters. For *S. aureus*, 49 % of the
162 genomes contain the c335-ermA cluster and in *S. pneumonia* 55 % contain the
163 c17-tetM cluster and 14 % contain the c310-ermB cluster. Furthermore, 88 % of
164 the *S. aureus* genomes carry the c5-mecA gene cluster responsible for the
165 methicillin resistance Staphylococcus aureus (MRSA) phenotype. *N. meningitidis,*
166 *N. gonorrhoeae, C. jejuni and S. pyogenes* all have fewer than 10 unique resistance
167 gene clusters suggesting that these species do not harbor many acquired
168 resistance genes compared to other pathogens (Table 1). Comfortingly, these
169 findings are in agreement with prior studies of abundant resistance genes in
170 different bacterial species[9,10,12,27,28]. Yet, in contrast to previous studies, we
171 employ a consistent analysis pipeline across all available genomes, which can be
172 readily updated with availability of additional genomic information.
173
174 In general there is a trend that resistance gene clusters with many hits are also
175 observed in more species (Figure 1). For instance, the c205-TEM beta-lactamase
176 gene cluster is by far the most disseminated resistance gene cluster as well as
177 one of the gene clusters with most hits. This gene cluster contains 155 unique
178 genes and is found in almost 2000 genomes across 12 of the 17 analyzed species
179 (un-clustered prevalence available on: www.resistome.biosustain.dtu.dk). Other
180 highly disseminated resistance gene clusters found in more than 8 species
181 include: the trimethoprim resistance gene cluster c449-dfrA, the streptomycin
182 resistance gene clusters c270-strA and c232-strB, the sulfamethoxazole
183 resistance gene clusters c146-sul and c215-sul, and the tetracycline resistance
184 gene clusters c17-tetM, c65-tetB and c67-tetA. In contrast, the c5-mecA gene
185 cluster is only found in *S. aureus*; yet, it is the most abundant resistance gene in
186 the whole dataset, being present in more than 4000 genomes reflecting the
187 selective sequencing of MRSA strains. Although the number and the clonality of
188 the sequenced genomes biases this analysis, it does highlight how some
189 resistance genes have become very successful in disseminating to many different
190 species.

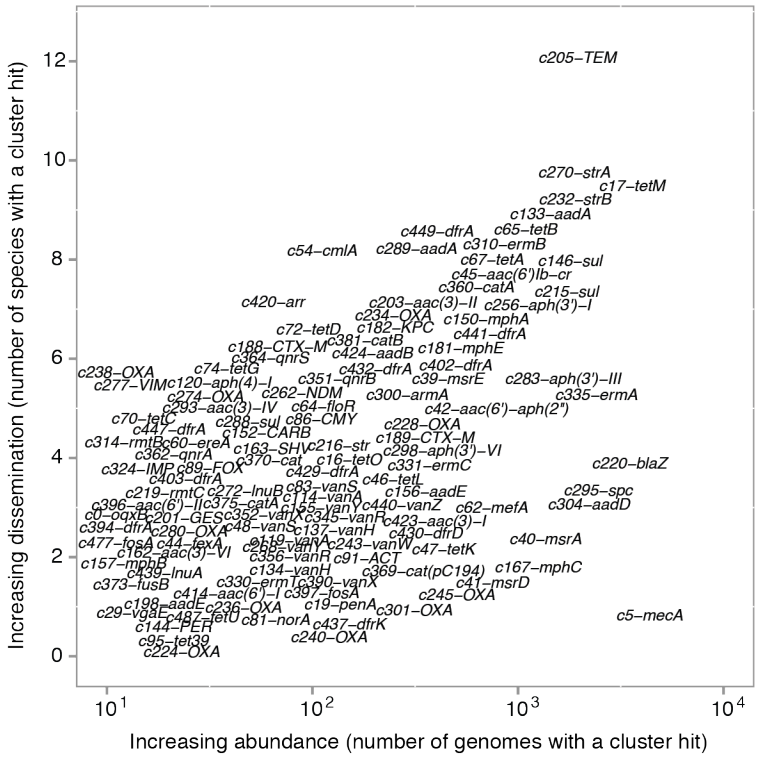


Figure 1. Resistance gene cluster dissemination as a function of abundance. Each of the 187 resistance gene clusters identified in the genomes isolated from humans are plotted according to how many of the 17 analyzed species they were found in and the total number of genomes they were found in. To reduce over-plotting, each data point has been jittered on both axes.

Functional distribution of gene clusters

In order to associate each resistance gene cluster with the antibiotic class that it likely confers resistance to and a mechanism by which it does so, we used ResFam to annotate the resistance gene clusters[29]. When stratified by antibiotic class, the abundances of the resistance gene clusters are differentially distributed across the 16 species (*M. tuberculosis* not included). Whereas gene clusters comprising beta-lactam and aminoglycoside resistance genes are found in genomes from most species, gene clusters comprising vancomycin and macrolide resistance genes are only abundant in a small subset of species (Figure 2a). Indeed, vancomycin resistance gene clusters are mainly found in *Enterococcus* spp. and macrolide resistance genes are predominantly found in

211 Gram-positive species. This likely reflects an antibiotic-dependent difference in
212 the dissemination of resistance genes, where the widely used beta-lactam and
213 aminoglycoside antibiotics select for a broad dissemination of beta-lactam and
214 aminoglycoside resistance genes while vancomycin and the macrolide antibiotics
215 are mainly used against *Enterococcus* spp. and *S. aureus* and Gram-positive
216 species, respectively (Figure 2a). Still, macrolide resistance genes do occur in
217 Gram-negative species, which is probably a result of co-selection with other
218 resistance genes.
219
220 When stratifying the resistance gene clusters into the different mechanisms
221 underlying the resistance there are also clear divisions between Gram-positive
222 and Gram-negative species. It is interesting to note that resistance gene clusters
223 that encode phosphotransferases, which are commonly involved in
224 aminoglycoside resistance, seem to be more abundant in Gram-negative species
225 than Gram-positive species. Instead, Gram-positive species appear to rely on
226 acetyltranseferases and nucleotidyltransferases to achieve aminoglycoside
227 resistance. Gene clusters encoding rRNA methyltransferases and protection
228 proteins, commonly conferring resistance to peptidyl transferase inhibitors and
229 tetracycline, respectively, are more abundant in Gram-positive species (Figure
230 2b). For the rRNA methyltransferases, this likely reflects the fact that that the
231 resistance conferred by this mechanism is to antibiotics such as macrolides and
232 streptogramins, both of which are mainly used to treat Gram-positive infections.
233 In the case of the protection proteins the effect of the mechanism is different.
234 This mechanism mainly gives resistance to tetracycline antibiotics, which are
235 effective against both Gram-positive and Gram-negative species. However, while
236 Gram-positive species achieve resistance via ribosomal protection proteins,
237 Gram-negative species generally rely on efflux pumps to clear the drug from the
238 cell[9]. This difference might be due to the different cell physiology disfavoring
239 efflux pumps in Gram-positive species. An exception to this division of
240 mechanism between Gram-positive and Gram-negative species is the Gram-
241 negative species *C. jejuni* that commonly has the tetracycline resistance gene *tetO*
242 encoding a ribosomal protection protein.
243

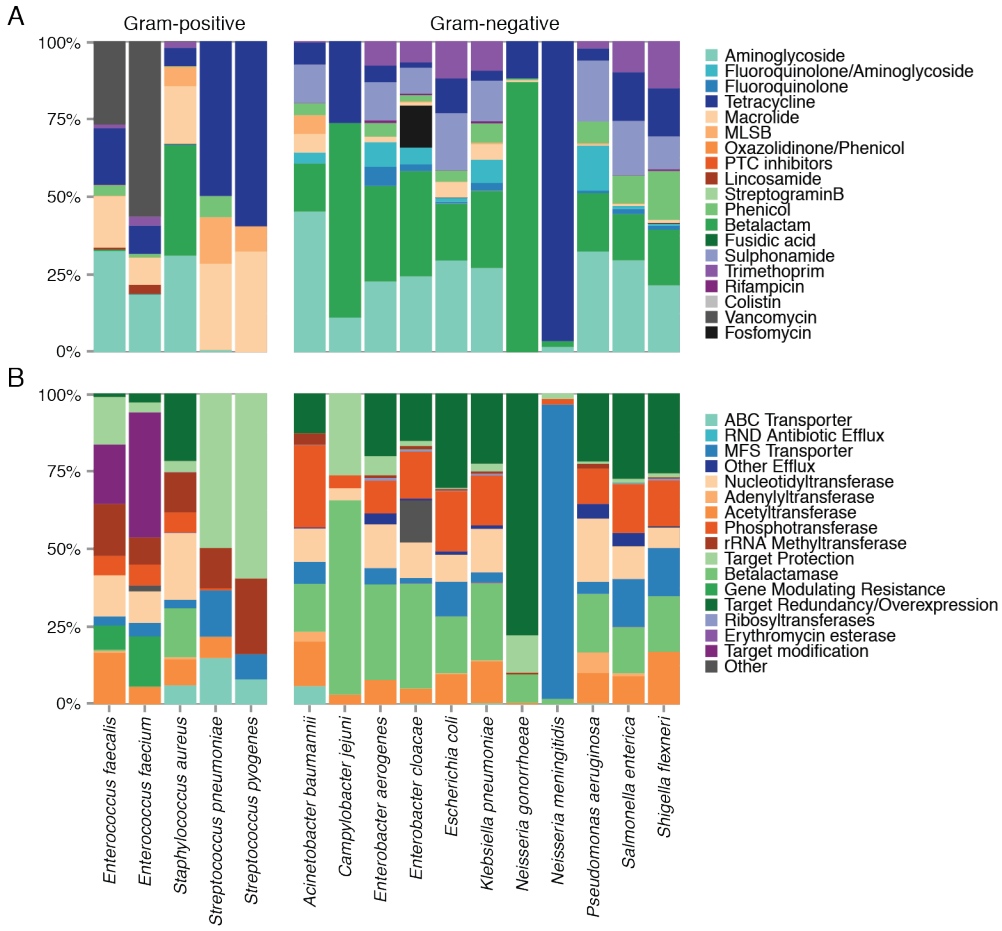


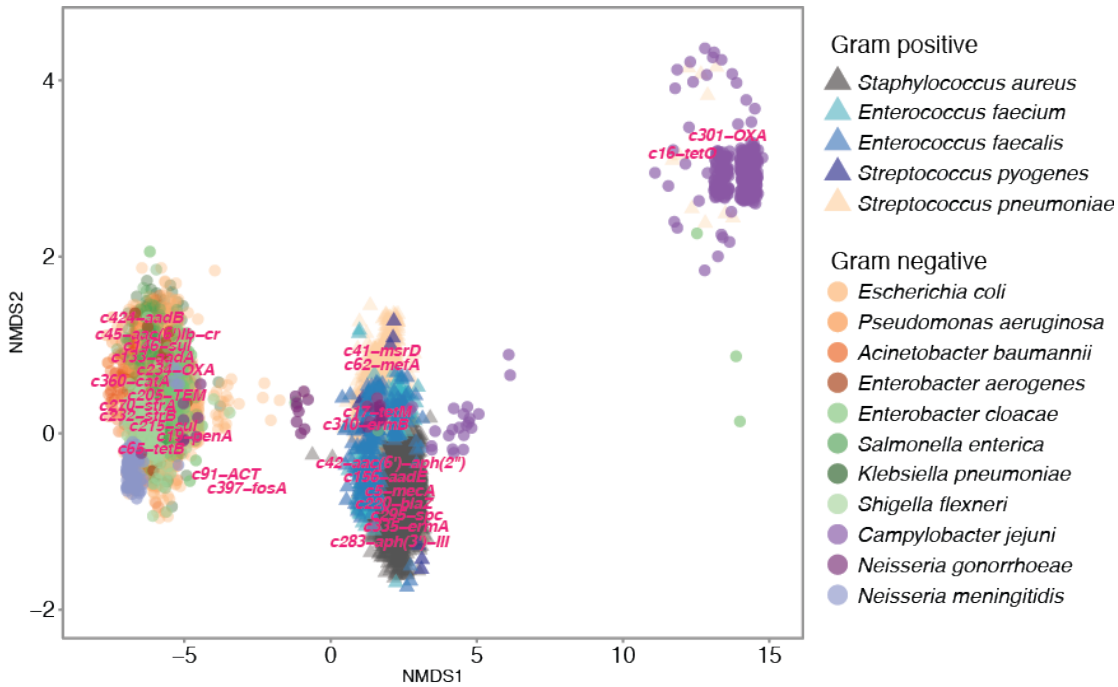
Figure 2. Overview of resistance mechanism and antibiotic category.
For all cluster hits across the genomes within a species, the relative distributions of the hits are shown according to the resistance mechanism (A) and the antibiotic category that the cluster confers resistance to (B). The mechanism category is assigned using Resfam and the drug category is implemented from the ResFinder metadata. (Abbreviations: MLSB macrolide, lincosamide, streptogramin B; PTC peptidyl transferase center; ABC ATP-binding cassette; RND resistance nodulation division; MFS major facilitator superfamily)

Genome level cluster profiles

In order to systematically analyze the distribution of resistance gene cluster hits across different species, we used non-metric multi-dimensional scaling (NMDS) analysis to identify genomes with similar cluster hit profiles (Figure 3). The analysis clearly reveals that Gram-positive and Gram-negative species generally have distinct cluster hit profiles, highlighting that the resistome is largely non-

261 overlapping between Gram-positive and Gram-negative species. The different
262 cluster hit profiles are most likely driven by multiple factors including
263 differences in cell physiology, antibiotic usage and ability to receive and maintain
264 mobile genetic elements.
265
266 By overlaying the most common resistance gene clusters, the analysis also
267 highlights associations between gene cluster and species. For instance, *S. aureus*
268 genomes are highly clustered by the c5-mecA and c220-blaZ gene clusters (88%
269 and 70 % abundance, respectively), the former causing the MRSA phenotype and
270 only found in *S. aureus*. The *Enterococcus faecalis* genomes cluster by the c310-
271 ermB and c17-tetM gene clusters (58 % and 52 % abundance, respectively),
272 while *E. faecium* is spread out between the c310-ermB and c156-aadE gene
273 clusters (76% and 67 % abundance, respectively). *Streptococcus pneumoniae*
274 forms two clusters, one driven by the c310-ermB and c17-tetM gene clusters (14
275 % and 55 % abundance, respectively) overlapping with *E. faecalis,* and one
276 driven by the c41-msrD and c62-mefA gene clusters (both 17 % abundance). In
277 contrast, the cluster hit profiles of Gram-negative genomes are more tightly
278 clustered around the highly disseminated c205-TEM, c146-sul, c215-sul, c270-
279 strA and c232-strB clusters, suggesting a larger extent of resistance gene
280 dissemination within Gram-negative species compared to Gram-positive. Only
281 the *N. meningitides* and a few of the *S. enterica* and *E. cloacae* genomes cluster
282 closely together, the remaining genomes make up one dense cluster with
283 multiple species. This reflects that a major overlap exists between the most
284 common resistance gene clusters across the Gram-negative species, which is also
285 reflected in Table 1 and Supplementary figure 1. Interestingly, *Campylobacter*
286 *jejuni* species constitute two clusters distinct from the other clusters. This
287 distinction is mainly driven by the c301-OXA gene cluster (68 % abundance),
288 which is unique to *C. jejuni*. In addition, the gene cluster c16-tetO (29 %
289 abundance) is also a distinct hallmark of *C. jejuni*. Interestingly, the gene clusters
290 c16-tetO, c283-aph(3')-III and c198-aadE (29%, 5% and 4 % abundance
291 respectively), commonly found in *C. jejuni*, are generally associated with Gram-
292 positive genomes and therefore make some *C. jejuni* genomes cluster with the
293 Gram-positive species[30,31]. Notably, the ability of *C. jejuni* to maintain

294 predominantly Gram-positive resistance genes may enable this species to be a
295 bridging species that can link the Gram+ and Gram- resistome. Likewise, for *N.*
296 *gonorrhoeae,* the predominantly Gram-positive c17-tetM gene cluster (7 %
297 abundance) drives the clustering towards the Gram-positive profiles[31].
298



299
300
301 Figure 3. NMDS of the genomes based on the resistance gene profile.
302 NMDS analysis based in the presence or absence of the resistance gene clusters
303 in the individual genomes. The genome observations are overlaid with the
304 cluster scores of the three most abundant clusters for each species. The plot is
305 based on binary Bray-Curtis distances, stress = 0.042. To reduce over-plotting,
306 each data point is jittered on both axes.
307
308 **Co-occurrence of resistance gene clusters**
309 As our analyses are performed at the genome level, we were able to identify
310 resistance genes that repeatedly co-occur in the genome of different strains.
311 Genes with strong co-occurrence could be genetically linked and disseminated
312 by a specific integron, transposon or plasmid. For 14 of the 17 species, we
313 generated co-occurrence matrices based on the pairwise co-occurrence of the
314 resistance gene clusters, the remaining three species did not have enough cluster

315 hits to be included in the analysis. We calculated the frequency of co-occurrence
316 as the number of pairwise co-occurrences relative to the total number of
317 occurrences of each gene. To reduce spurious findings we limited the analysis to
318 genes that were found in more than 5 % of the genomes within a given species.
319 The resulting co-occurrence matrix gives information on the frequency by which
320 a given gene co-occurs with another gene (Supplementary figure 2, see
321 www.resistome.biosustain.dtu.dk). If two genes are always found together, they
322 have complete linkage. If a gene is observed together with another gene in 50 %
323 of the cases, it displays partial linkage and finally, when two genes are never
324 observed together, they display no linkage. It should be noted that, for a gene-
325 pair, A and B, it is possible for gene A to be completely linked to gene B, while
326 gene B is only partially linked to gene A. This would happen if gene B were found
327 in multiple genetically different contexts while gene A was only found in one
328 genetic context, i.e. a specific transposon. While a strong linkage between two or
329 more genes might indicate linked dissemination, it could also result from
330 sequencing of clonal lineages and thus simply reflect the dissemination of a
331 strain rather than a genetic element.
332
333 The extent of co-occurrences varies greatly from species to species. *E. coli, A.*
334 *baumannii, K. pneumoniae, S. enterica, S. flexneri, E. cloacae, E. faecalis, E. faecium*
335 and *S. aureus* have a large co-occurrence network, while *C. jejuni, E. aerogenes, N.*
336 *gonorrhoeae, P. aeruginosa* and *S. pneumoniae* have a small co-occurrence
337 network with just a few genes (Supplementary figure 2, see
338 www.resistome.biosustain.dtu.dk). *S. pyogenes, N. meningitidis* and *M.*
339 *tuberculosis* do not have enough gene clusters above the threshold and therefore
340 the co-occurrence network could not be computed.
341
342 For some gene clusters, the co-occurrence is well understood e.g. the co-
343 occurrence of *strA* and *strB,* which combined give high level streptomycin
344 resistance[32]. Similarly, the different *van* gene clusters conferring vancomycin
345 resistance in enterococci are strongly co-occurring, as resistance is achieved
346 through the concerted action of these genes[10].
347

348 The co-occurrence analysis also recapitulates known co-dissemination patterns
349 driven by linkage on mobile genetic elements. For instance, the macrolide
350 resistance gene cluster c335-ermA in *S. aureus* is located just upstream of the
351 spectinomycin resistance gene cluster c295-spc (Figure 4A). The two genes are a
352 part of transposon Tn554, which in turn is a part of a larger pathogenicity island
353 including the c5-mecA and the c304-aadD clusters[33,34]. As a result the clusters
354 c295-spc, c335-ermA and c304-aadD are all strongly linked to each other as well
355 as to c5-mecA; however, c5-mecA is found in many other contexts and therefore
356 not strongly linked to the three other resistance genes (Figure 4A).
357
358 In *E. coli* there is a strong linkage between the streptomycin/spectinomycin
359 resistance gene cluster c289-aadA and the trimethoprim resistance cluster c402-
360 dfrA, and a further linkage to the sulfamethoxazole resistance gene cluster c146-
361 sul. These resistance genes are commonly found together in class 1 integrons in
362 strains of both human and animal origin[35] (Figure 4B). Interestingly, of the 284
363 genomes where clusters c289-aadA and c402-dfrA co-occur, 242 (85 %) are
364 found in an integron with an IS26 insertion truncating the *intI1* gene (Figure 4B).
365
366

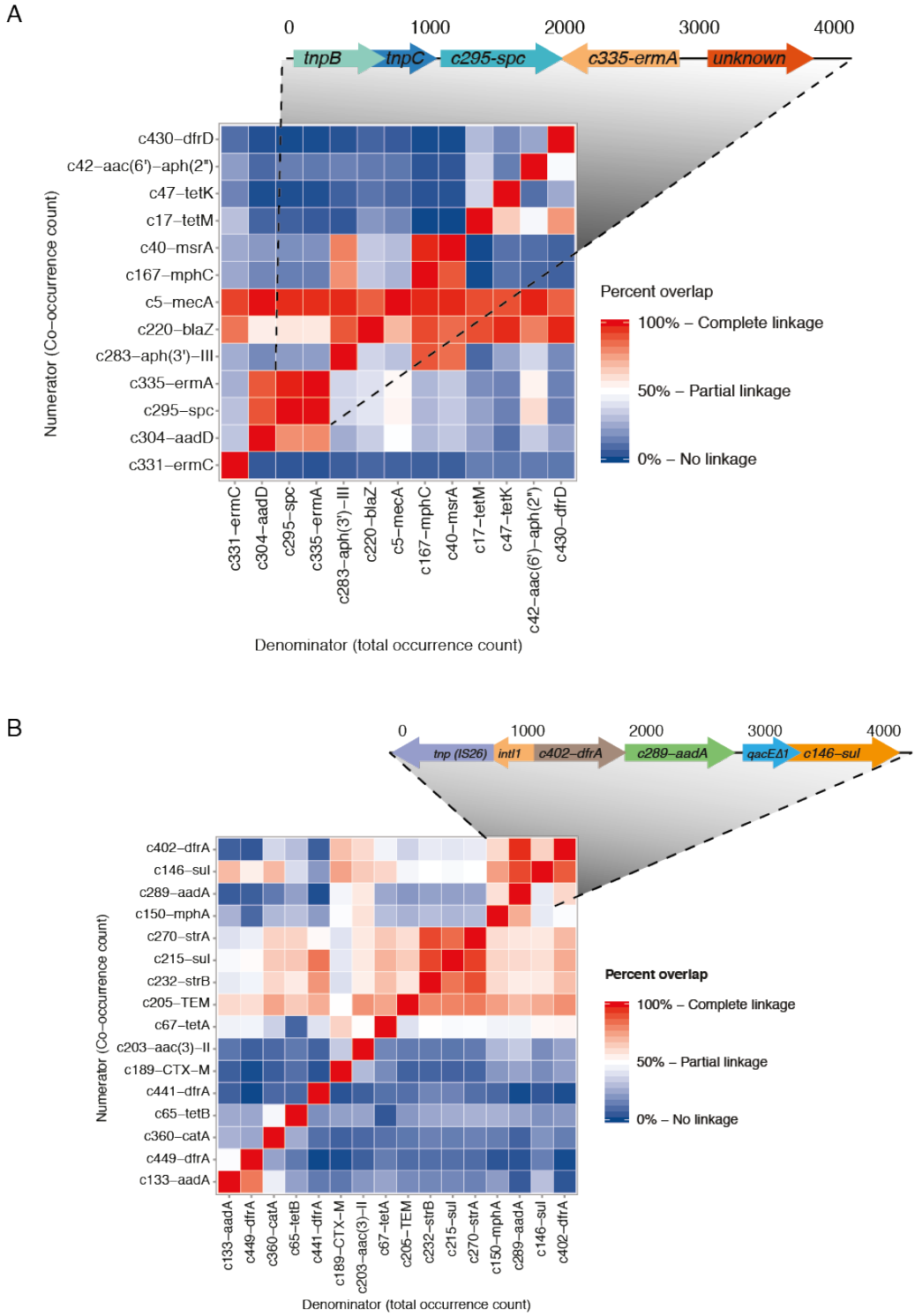


Figure 4. Co-occurrence matrix.

A) Co-occurrence matrix for S. aureus, the macrolide resistance gene c335-ermA is found in 2393 (49 %) of the *S. aureus* genomes and in 2382 (99.5 %) of the cases it is found together with the c295-spc gene, most commonly in the configuration depicted in the figure as part of the transposon Tn554, where the two genes are located next to each other. Co-occurrences are only calculated for

genes present in minimum 5 % of the analyzed genomes. B) Co-occurrence matrix for *E. coli* the aminoglycoside resistance gene cluster c289-*aadA* is found in 284 (12 %) of the *E. coli* genomes and in 242 (85 %) of the cases it is found together with the c402-*dfrA* gene cluster, most commonly in the configuration depicted in the figure as a class 1 integron with an IS26 insertion truncating the *intI1* gene. Co-occurrences are only calculated for genes present in minimum 5 % of the analyzed genomes.

In addition to identifying linked resistance genes, the co-occurrence analysis also identifies genes that are rarely or never found together. This can be caused by resistance genes being mobilized on different plasmids belonging to the same incompatibility group or by a high fitness cost of simultaneous carriage of two genes. For instance, hits to cluster c65-tetB are not found to co-occur with hits to cluster c67-tetA. A possible explanation for this observation could be a lack of selective benefit from simultaneous carriage of *tetA* and *tetB* due to their identical phenotype. The same phenomenon is observed for the three *dfrA* clusters c402, c441 and c449 (Figure 4B).

407 **Discussion**

408 The declining cost of whole-genome sequencing enables detailed studies of the
409 evolution and dissemination of pathogenic bacteria. Here, we present the results
410 of a genome-wide identification of resistance genes in publicly available
411 genomes of key pathogens isolated from humans. We have identified close
412 homologues of the main genetic determinants of antibiotic resistance in 17
413 common human pathogens. We found that 38% of the resistance gene clusters
414 from the query database were found in bacterial pathogens isolated from
415 humans, highlighting that most resistance gene clusters were not present in the
416 sequenced genomes.

417

418 As our analysis is done *in silico* with no biochemical validation we used stringent
419 thresholds requiring a hit to have a minimum of 95% sequence identity and a
420 minimum of 90% coverage to a query. This ensures, that hits have a high
421 likelihood of being true resistance genes. However, although our analysis
422 identifies close homologues of resistance genes it does not necessarily imply
423 phenotypic resistance, as genes may not always be expressed.
424 To further ensure a reliable prediction of resistance genes we exclusively
425 focused on non-mutational resistance, although we acknowledge that clinically
426 important resistances also occur via mutations in native genes. In addition, some
427 species may be intrinsically resistant due to native genes conferring resistance,
428 and while this is an important phenomenon such intrinsic resistance
429 mechanisms are not included in our study.

430

431 Importantly, our analysis does not represent a study of horizontal gene transfer.
432 We used the ResFinder database as a reference database for resistance genes as
433 this is a manually curated and actively maintained database of acquired
434 resistance genes[24]. Other studies have focused on large-scale identification of
435 HGT typically requiring computational intensive methods [6,36].

436

437 In addition to the Resfinder database other databases of resistance genes, with
438 varying degrees of curation and maintenance, exist[37]. Currently, the
439 Comprehensive Antibiotic Resistance Database (CARD) is the largest database

440 including resistance-conferring variants of native genes. However, when
441 comparing the CARD subset of acquired resistance genes to the ResFinder
442 database, we found that CARD contained many native genes, for instance 12% of
443 the genes that were unique to CARD compared to ResFinder (representing 43
444 gene clusters) were native to *E. coli* K12, which is broadly considered to be a
445 sensitive organism. Such genes would lead to overestimation of resistance gene
446 abundance in our analysis (see Supplementary table 4 for a complete list of
447 CARD exclusive genes). In contrast, just 16 clusters from the ResFinder database
448 were identified as native resistance genes.
449
450 Currently, many of the available genomes represent sequencing efforts directed
451 towards specific phenotypes such as MRSA or carbapenem-resistant *K.*
452 *pneumonia.* While these highly resistant pathogens are a major threat to human
453 health, their overrepresentation in publicly available databases limit studies into
454 the general trends of antibiotic resistance evolution and spread. To obtain a
455 more unbiased view of the global emergence and spread of resistance genes,
456 whole genome sequences from organisms isolated as a part of routine
457 diagnostics in medical institutions around the world should be deposited to
458 public databases. Alternatively, sequencing of key pathogens, representatively
459 sampled at regular intervals, could provide valuable data on the trends of
460 resistance gene evolution and dissemination.
461
462 As studies have found that resistance genes can be found in all environments it is
463 relevant to identify the genes that emerge in common human pathogens[1]. With
464 our generalized analysis we have generated a subset of resistance genes found in
465 common pathogens of human origin. This database represents the current subset
466 of resistance genes that are particular relevant to human health. We believe that
467 the database will be useful in assessing the interaction between environmental
468 and clinical resistomes based on metagenomic and genome sequencing studies.
469 Further updates of the database will identify the movement of genes from the
470 general resistome into the clinically relevant resistome. In turn, this can be used
471 to rank the risk-potential of different reservoirs of antibiotic resistance genes[38].
472 Furthermore, if representative collections of bacterial genomes from clinical

473 microbiological departments around the globe were available for analysis,
474 approaches such as the one presented here could be used to more effectively
475 monitor the changing patterns of antibiotic resistance genes and identify the
476 genetics determinants that contribute most to the dissemination of antibiotic
477 resistance.
478
479
480 **Materials and methods**
481
482 Genome database.
483 All available assembled refseq genomes for the investigated species were
484 downloaded from the National Center for Biotechnology Information (NCBI)
485 (www.ncbi.nlm.nih.gov/refseq/) (downloaded July 2016). Bacterial isolates from
486 humans were identified by searching for the term “human” or “homo” in the host
487 field of the genbank file.
488
489 Identifying resistance genes in the human isolates.
490 Using BLAST, we searched the downloaded genomes for the presence of each
491 gene from the Resfinder resistance gene database (downloaded July 2016)[24]. In
492 the BLAST command the following parameters were used: -perc_identity 95 -
493 max_target_seqs 500000000 -task megablast -outfmt '6 std sstrand qlen slen
494 sseq'. The output file was filtered to only include hits with query coverage of >=
495 90%, query coverage was calculated as length/qlen. An overview of the data
496 analysis is given in Supplementary figure S4, at
497 www.resistome.biosustain.dtu.dk.
498
499 Clustering of hits.
500 In order to present the results in a more concise format, the hits were binned
501 into clusters. The resistance gene clusters were based on a clustering of the
502 resistance gene database. The database was clustered using CDhit (80 % identity,
503 80 % coverage)[39]. Using the cluster information from the resistance gene
504 database, the BLAST hits were binned into resistance gene clusters. Each
505 resistance gene cluster was represented by the most common sequence.

506

507 Systematic annotation of the resistance gene clusters.

508 In order to systematically annotate the resistance gene clusters with mechanism
509 of resistance we used the hmm database from ResFam[29]. To ensure reliable
510 annotation we first predicted the open reading frame (ORF) in the
511 representative sequence using GeneMarks.hmm[40]. Subsequently the mechanism
512 of resistance for the representative protein sequence was annotated with
513 HMMER (hmmscan) using the ResFam database (Resfams-full.hmm)[29,41].

514

515 Calculation of resistance cluster frequency.

516 For each bacterial species, the frequency of each resistance gene or resistance
517 gene cluster was calculated. This resulted in a table of cluster abundances per
518 species. This table was manually curated to remove genes belonging to the core
519 genome. Genes with high abundance (>93 %) that could be identified as
520 endogenous based on a literature search were removed (see supplementary
521 table 3 for the complete cluster list along with references for removing). The
522 resulting table represents the resistance genes found in common human
523 pathogens referred to as the clinical resistome.

524

525 Generation of the un-clustered database.

526 As the query database contained many highly similar sequences, e.g. single
527 nucleotide variants of the same resistance gene, there were often multiple hits to
528 the same subject. In these cases, the bitscore was used to identify the best hit. In
529 addition, as the query database contained highly similar genes with different
530 lengths, e.g. same resistance gene but with an alternative start codon, it was
531 possible for a subject region to have two high scoring hits. To overcome
532 overestimation of resistance gene abundances associated with these variations,
533 hits were grouped according to their location on the target sequence, such that if
534 two hits had start and stop positions within 20 bp. of each other, respectively,
535 only the longest hit was considered.

536 Data presentation.

537 All data analysis was done in R, using the packages ggplot2 and vegan[42,43]. For
538 the NMDS analysis, Bray-Curtis distances were calculated with binary = TRUE.

539 For the co-occurrence analysis only resistance gene clusters that had >5%
540 abundance were included. For each pair of resistance gene clusters both
541 frequencies of co-occurrence were calculated, i.e. for clusters A and B, both the
542 frequencies of A+B/A and A+B/B were calculated.
543
544 Flanking region analysis.
545 To confirm genetic linked dissemination, BLAST was used to identify the
546 genomic location of the resistance gene using the subject sequence (sseq) as
547 query. Subsequently, the flanking regions 2 kb up- and downstream from the
548 gene were extracted. The regions were clustered using CD-hit (90 % identity, 90
549 % coverage). Next, the ORFs in the extracted regions were identified using
550 Genemarks.hmm and clustered into gene families CD-hit (95 % identity, 95 %
551 coverage) and annotated manually. Only the most common region signature is
552 shown in figures 4 and 5.
553
554 Data availability.
555 A complete overview of all resistance genes at both single gene (un-clustered)
556 and cluster level for each species is available at the web site
557 www.resistome.biosustain.dtu.dk. In addition, all supplementary information,
558 including fasta files with all un-clustered resistance genes, is available from the
559 same web site.
560
561
562 **References:**
563

564 1. Allen, H. K. *et al.* Call of the wild: antibiotic resistance genes in natural
565 environments. *Nature Reviews Microbiology* **8,** 251–259 (2010).
566 2. van Schaik, W. The human gut resistome. *Philos. Trans. R. Soc. Lond., B, Biol.*
567 *Sci.* **370,** 20140087–20140087 (2015).
568 3. D'Costa, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477,** 457–461
569 (2011).
570 4. Bhullar, K. *et al.* Antibiotic resistance is prevalent in an isolated cave
571 microbiome. *PLoS ONE* **7,** e34953 (2012).
572 5. Wintersdorff, von, C. J. H. *et al.* Dissemination of Antimicrobial Resistance
573 in Microbial Ecosystems through Horizontal Gene Transfer. *Front.*
574 *Microbio.* **7,** 173 (2016).
575 6. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange

576 connecting the human microbiome. *Nature* **480,** 241–244 (2011).
577 7. Dantas, G. & Sommer, M. O. Context matters — the complex interplay
578 between resistome genotypes and resistance phenotypes. *Current Opinion*
579 *in Microbiology* **15,** 577–582 (2012).
580 8. Munck, C. *et al.* Limited dissemination of the wastewater treatment plant
581 core resistome. *Nature Communications* **6,** 8452 (2015).
582 9. Roberts, M. C. Update on acquired tetracycline resistance genes. *FEMS*
583 *Microbiology Letters* **245,** 195–203 (2005).
584 10. Courvalin, P. Vancomycin resistance in gram-positive cocci. *Clin. Infect. Dis.*
585 **42 Suppl 1,** S25–34 (2006).
586 11. Munoz-Price, L. S. *et al.* Clinical epidemiology of the global expansion of
587 Klebsiella pneumoniae carbapenemases. *The Lancet Infectious Diseases* **13,**
588 785–796 (2013).
589 12. Bush, K. Alarming β-lactamase-mediated resistance in multidrug-resistant
590 Enterobacteriaceae. *Current Opinion in Microbiology* **13,** 558–564 (2010).
591 13. Tangden, T. & Giske, C. G. Global dissemination of extensively drug-
592 resistant carbapenemase-producing Enterobacteriaceae: clinical
593 perspectives on detection, treatment and infection control. *J. Intern. Med.*
594 **277,** 501–512 (2015).
595 14. McAdam, P. R., Richardson, E. J. & Fitzgerald, J. R. High-throughput
596 sequencing for the study of bacterial pathogen biology. *Current Opinion in*
597 *Microbiology* **19,** 106–113 (2014).
598 15. Sintchenko, V. & Holmes, E. C. The role of pathogen genomics in assessing
599 disease transmission. *BMJ* **350,** h1314–h1314 (2015).
600 16. Tyson, G. H. *et al.* WGS accurately predicts antimicrobial resistance in
601 Escherichia coli. *Journal of Antimicrobial Chemotherapy* **70,** 2763–2769
602 (2015).
603 17. Gordon, N. C. *et al.* Prediction of Staphylococcus aureus Antimicrobial
604 Resistance by Whole-Genome Sequencing. *Journal of Clinical Microbiology*
605 **52,** 1182–1191 (2014).
606 18. Stoesser, N. *et al.* Predicting antimicrobial susceptibilities for Escherichia
607 coli and Klebsiella pneumoniae isolates using whole genomic sequence
608 data. *Journal of Antimicrobial Chemotherapy* (2013).
609 19. Zhao, S. *et al.* Whole-Genome Sequencing Analysis Accurately Predicts
610 Antimicrobial Resistance Phenotypes in Campylobacter spp. *Applied and*
611 *Environmental Microbiology* **82,** 459–466 (2016).
612 20. Walker, T. M. *et al.* Whole-genome sequencing for prediction of
613 Mycobacterium tuberculosis drug susceptibility and resistance: a
614 retrospective cohort study. *The Lancet Infectious Diseases* **15,** 1193–1202
615 (2015).
616 21. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome
617 sequence data for Staphylococcus aureus and Mycobacterium tuberculosis.
618 *Nature Communications* **6,** 10063–14 (2015).
619 22. Rice, L. B. Federal funding for the study of antimicrobial resistance in
620 nosocomial pathogens: no ESKAPE. *J. Infect. Dis.* **197,** 1079–1081 (2008).
621 23. Boucher, H. W. *et al.* Bad bugs, no drugs: no ESKAPE! An update from the
622 Infectious Diseases Society of America. *Clin. Infect. Dis.* **48,** 1–12 (2009).
623 24. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J.*
624 *Antimicrob. Chemother.* **67,** 2640–2644 (2012).

625 25. Héritier, C. *et al.* Characterization of the naturally occurring oxacillinase of
626 Acinetobacter baumannii. *Antimicrob. Agents Chemother.* **49,** 4174–4179
627 (2005).
628 26. Almeida Da Silva, P. E. A. & Palomino, J. C. Molecular basis and mechanisms
629 of drug resistance in Mycobacterium tuberculosis: classical and new drugs.
630 *J. Antimicrob. Chemother.* **66,** 1417–1430 (2011).
631 27. Choi, T., Yoo, K. H. & Lee, S.-J. Changing Epidemiology of Extended
632 Spectrum Beta-Lactamases Pathogen of Urinary Tract. *Urogenital Tract*
633 *Infection* **10,** 74 (2015).
634 28. Grundmann, H. *et al.* Occurrence of carbapenemase-producing Klebsiella
635 pneumoniae and Escherichia coli in the European survey of
636 carbapenemase-producing Enterobacteriaceae (EuSCAPE): a prospective,
637 multinational study. *The Lancet Infectious Diseases* **17,** 153–163 (2017).
638 29. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic
639 resistance determinants reveals microbial resistomes cluster by ecology.
640 **9,** 207–216 (2014).
641 30. Qin, S. *et al.* Identification of a novel genomic island conferring resistance
642 to multiple aminoglycoside antibiotics in Campylobacter coli. *Antimicrob.*
643 *Agents Chemother.* **56,** 5332–5339 (2012).
644 31. Roberts, M. C. in *Antibiotic Discovery and Development* 543–568 (Springer
645 US, 2011). doi:10.1007/978-1-4614-1400-1_16
646 32. Sundin, G. W. & Bender, C. L. Dissemination of the strA-strB streptomycin-
647 resistance genes among commensal and pathogenic bacteria from humans,
648 animals, and plants. *Mol. Ecol.* **5,** 133–143 (1996).
649 33. Murphy, E., Huwyler, L. & de Freire Bastos, M. D. C. Transposon Tn554:
650 complete nucleotide sequence and isolation of transposition-defective and
651 antibiotic-sensitive mutants. *EMBO J.* **4,** 3357–3365 (1985).
652 34. Appelbaum, P. C. Microbiology of antibiotic resistance in Staphylococcus
653 aureus. *Clin. Infect. Dis.* **45 Suppl 3,** S165–70 (2007).
654 35. Fluit, A. C. & Schmitz, F. J. Class 1 integrons, gene cassettes, mobility, and
655 epidemiology. *European Journal of Clinical Microbiology & Infectious*
656 *Diseases* **18,** 761–770 (1999).
657 36. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured
658 from global to individual scales. *Nature* **535,** 435–439 (2016).
659 37. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the
660 comprehensive antibiotic resistance database. *Nucleic Acids Research* **45,**
661 D566–D573 (2017).
662 38. Martínez, J. L., Coque, T. M. & Baquero, F. What is a resistance gene?
663 Ranking risk in resistomes. *Nature Reviews Microbiology* **13,** 116–123
664 (2014).
665 39. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing
666 large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–
667 1659 (2006).
668 40. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene
669 finding. *Nucleic Acids Research* **26,** 1107–1115 (1998).
670 41. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763
671 (1998).
672 42. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer New
673 York, 2009). doi:10.1007/978-0-387-98141-3

43. Oksanen, J. Vegan: ecological diversity. 1–11 (2013).

Acknowledgements

This research was funded by the EU H2020 ERC-20104-STG LimitMDR (638902) and the Danish Council for Independent Research Sapere Aude programme DFF - 4004-00213. MOAS acknowledges additional funding from the Novo Nordisk Foundation and The Lundbeck Foundation.