UNIVERSITYOF BIRMINGHAM

Prediction of the intestinal resistome by a three-dimensional structure-based method

Ruppé, Etienne; Ghozlane, Amine; Tap, Julien; Pons, Nicolas; Alvarez, Anne-Sophie; Maziers, Nicolas; Cuesta, Trinidad; Hernando-Amado, Sara; Clares, Irene; Martínez, Jose Luís; Coque, Teresa M; Baquero, Fernando; Lanza, Val F; Máiz, Luis; Goulenok, Tiphaine; de Lastours, Victoire; Amor, Nawal; Fantin, Bruno; Wieder, Ingrid; Andremont, Antoine

Document Version
Peer reviewed version

Link to publication on Research at Birmingham portal

1   **Prediction of the intestinal resistome by a 3D structure-based method**

2   Etienne Ruppé* (1, 2), Amine Ghozlane* (1, 3, 4) Julien Tap*§ (1), Nicolas Pons (1), Anne-Sophie

3   Alvarez (1), Nicolas Maziers (1), Trinidad Cuesta (5), Sara Hernando-Amado (5), Irene Clares (5), Jose

4   Luís Martínez (5), Teresa M. Coque (6, 7, 8), Fernando Baquero (6, 7, 8), Val F. Lanza (6, 7), Luis Maiz

5   (9), Tiphaine Goulenok (10), Victoire de Lastours (2, 10), Nawal Amor (10), Bruno Fantin (2, 10), Ingrid

6   Wieder (11), Antoine Andremont (2, 11), Willem van Schaik (12, 13), Malbert Rogers (12), Xinglin Zhang

7   (12), Rob J.L. Willems (12), Alexandre G. de Brevern (14), Jean-Michel Batto (1), Hervé M. Blottière (1),

8   Pierre Léonard (1), Véronique Léjard (1), Aline Letur (1), Florence Levenez (1), Kevin Weiszer (1),

9   Florence Haimet (1), Joël Doré (1), Sean P. Kennedy (1, 4), S. Dusko Ehrlich (1, 15)

10

11   (1) MGP MetaGénoPolis, INRA, Université Paris-Saclay, 78350 Jouy en Josas, France

12   (2) IAME, UMR 1137, INSERM, Paris Diderot University, Sorbonne Paris Cité,

13   (3) Bioinformatics and Biostatistics Hub, C3BI, Institut Pasteur, USR 3756 IP CNRS, Paris, France.

14   (4) Biomics, CITECH, Institut Pasteur, Paris, France.

15   (5) Centro Nacional de Biotecnología, CSIC, Madrid, Spain.

16   (6) Servicio de Microbiología. Instituto, Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid,

17   Spain

18   (7) CIBER en Epidemiología y Salud Pública (CIBER-ESP), Madrid, Spain

19   (8) Unidad de Resistencia a Antibióticos y Virulencia Bacteriana (RYC-CSIC), Madrid, Spain.

20   (9) Unit for Cystic Fibrosis, Ramon y Cajal University Hospital, Madrid, Spain

21   (10) Internal Medicine Department, Beaujon Hospital, AP-HP, Clichy, France

22   (11) Bacteriology Laboratory, Bichat-Claude Bernard Hospital, AP-HP, Paris, France

23   (12) Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, the Netherlands

24   (13) Institute of Microbiology and Infection, University of Birmingham, Edgbaston, Birmingham B15 2TT,

25   United Kingdom

26   (14) INSERM UMR_S 1134, Paris Diderot University, Sorbonne Paris Cité, Université de la

27   Réunion, Université des Antilles, INTS, GR-Ex, Paris, France

28   (15) Centre of Host Microbiome Interactions, King's college, London, United Kingdom

29

30   *The authors equally contributed to the study

31

32   § Current affiliation: Danone Nutricia Research, Palaiseau – France

33

34   **Corresponding author**

35   Etienne RUPPE (PharmD, PhD)

36   Laboratoire de Bactériologie

37   Hôpital Bichat-Claude Bernard

38   46 rue Henri Huchard

39   75018 Paris

40   France

41   Phone: +33(0) 1 40 25 85 04

42   Fax: +33(0) 1 40 25 85 81

43   etienne.ruppe@inserm.fr

44

Opening paragraph

The intestinal microbiota is considered to be a major reservoir of antibiotic resistance determinants (ARDs) that could potentially be transferred to bacterial pathogens via mobile genetic elements (MGEs). Yet, this assumption is poorly supported by empirical evidence due to the distant homologies between known ARDs (mostly from culturable bacteria) and ARDs from the intestinal microbiota. Consequently, an accurate census of intestinal ARDs (*i.e.* the intestinal resistome) has not yet been fully determined. For this purpose, we developed and validated an annotation method (called pairwise comparative modelling, PCM) based on 3D structure (homology comparative modelling) leading to the prediction of 6,095 ARDs in a catalogue of 3.9 million proteins from the human intestinal microbiota. We found that the majority of predicted ARDs (pdARDs) were distantly related to known ARDs (mean amino acid identity 29.8%) and found evidence supporting their transfer between species. According to the composition of their resistome, we were able to cluster subjects from the MetaHIT cohort (n=663) into 6 "resistotypes" that were connected to the previously described enterotypes. Finally, we found that the relative abundance of pdARDs was positively associated with gene richness, but not when subjects were exposed to antibiotics. Altogether, our results indicate that the majority of intestinal microbiota ARDs can be considered as intrinsic to the dominant commensal microbiota and that these genes are rarely shared with bacterial pathogens.

Introduction

65 Antimicrobial resistance is one of the major threats to health identified by the World Health Organization

66 for the next decades. The intestinal microbiota plays a pivotal role in this phenomenon as it harbours a

67 vast diversity of bacterial species, some of them possessing antibiotic resistance determinants (ARDs)

68 that may enable their survival under antibiotic exposure. Previous studies attempted to identify ARDs in

69 the intestinal microbiota[2–4] but were confounded by the distant homologies between known ARDs

70 (mostly from culturable bacteria) and ARDs from the intestinal microbiota (which are generally not

71 cultured)[5,6]. For these reasons, bioinformatic tools based on sequence comparison (ARG-ANNOT[7],

72 CARD – RGI[8], Resfinder[9], DeepARG[10]) or motif detection (Resfams[11]) are often unsuccessful in

73 characterising the diversity of ARDs from metagenomic datasets. Indeed, there is no consensus on an

74 optimal approach to detect ARDs in metagenomic datasets. Consequently, an accurate census of

75 intestinal ARDs (*i.e.* the intestinal resistome) has not yet been fully determined.

76 While many bacteria have intrinsic, chromosomally-encoded ARDs and the capability of increasing

77 resistance through mutation, they can also enrich their resistance capabilities through the acquisition of

78 exogenous ARDs located on mobile genetic elements (MGEs) such as plasmids, transposons or

79 phages. The intestinal microbiota harbours thousands of bacterial species including well-known

80 pathogens (e.g. *Enterobacteriaceae* and *Enterococcus spp.*). This unique environment is assumed to

81 be a reservoir of ARDs that can potentially be transferred to bacterial pathogens[13]. Nonetheless despite

82 the high selective pressure exerted on the intestinal microbiota by over seven decades of intensive

83 antibiotic usage, a very low number of transfer events from an intestinal commensal to a bacterial

84 pathogen have been observed[14,15]. This challenges the hypothesis of a mobile resistome and the

85 assumption that the intestinal microbiota serves as a reservoir of ARDs to which pathogenic bacteria

86 have easy access[16]. In this study, our objective was to perform an extensive characterization of the

87 human gut resistome (including the capacity of ARDs to transfer between species) and to assess its

88 dynamics under various antibiotic exposures.

89

90 **Prediction of ARDs in the intestinal microbiota**

91 To predict ARDs in the intestinal microbiota, we developed a method based on protein homology

92 modelling (see methods) that we termed PCM (for "pairwise comparative modelling"). PCM is a generic

93 method using homology modelling to increase the specificity of functional prediction of proteins,

94 especially when they are distantly related to potential homologs. PCM uses a list of reference proteins

95 sequences from a given family, the ARD structures of this family (used as structural templates in protein

96 data bank [PDB] format) and a series of negative references (Figure 1A, Supplementary Figure 1, 2 and

97 3). Structural models are built using both the ARD reference and negative reference templates. Scores

98 generated from both positive and negative references are used to determine which model performed

99 the best. This is done using a machine-learning algorithm trained on 662 ARD and 522 negative

100 references. The PCM score equals the number of times the query was classified as an ARD for the

101 bootstraps performed, expressed as a percentage. Candidates with a PCM score ≥50% and an

102 alignment score with the reference template (TM score given by TM-align) ≥0.5[17] were predicted as

103 ARDs.

104 The performance of PCM to predict ARDs was assessed using *in vitro* and *in silico* methods. We

105 synthesized 71 candidate ARDs from 12 families (Table 1), and expressed them in *Escherichia coli* (see

106 methods). All 12 pdARDs sharing an amino acid identity >95% with a known ARDs had a detectable

107 resistance activity against antibiotics (Figure 1B). Resistance activity was also detected in 35/41 (85.3%)

108 of the predictions made with a good level of confidence (PCM score >99%, Tm score TmAlign>0.9) and

109 in 8/18 (44.4%) of the predictions with a lower level of confidence (PCM score <80%, Tm score

110 TmAlign<0.8). The mean amino acid identity of the functional pdARDs (good and fair predictions, n=43)

111 with known ARDs was 28.6% (range 19.4%-82.6%, Supplementary Table 1). We then tested PCM

112 against an experimentally-validated functional metagenomics dataset from soils[18]. In this case, PCM

113 was able to accurately identify 1,374 ARDs out of 1,423 hits (sensitivity 96.6%) (see methods). Finally,

114 we assessed the performances of PCM with incomplete proteins as inputs, and showed that PCM could

115 correctly predict ARDs when the available amino acid sequence was at least 40% complete

116 (Supplementary Figure 4). After the *in vitro* and *in silico* validation of the method, we used PCM to search

117 for ARDs in the in a catalogue made of 3,871,657 proteins which was built from the sequencing of faecal

118 samples of 396 human individuals (177 Danes and 219 Spanish) recruited in the MetaHIT project[19]. In

119 total, we predicted 6,095 ARDs (0.2% of the catalogue) from 20 ARD classes conferring resistance to

120 nine major antibiotic families[20]: beta-lactams (class A, B1-B2, B3, C and D beta-lactamases),

121 aminoglycosides (AAC(2'), AAC(3)-I, AAC(3)-II, AAC(6'), ANT, APH, RNA methylases), tetracyclines

122 (Tet(M), Tet(X)), quinolones (Qnr), sulphonamides (Sul), trimethoprim (DfrA), fosfomycin (Fos) and

123 glycopeptides (Van ligases) (Table 1 and Supplementary Table 1). With the same, extensively curated

124   reference ARDs census as input, only 67 ARDs would have been predicted according to conventional

125   BLASTP[21] search with a specific identity threshold (80% over 80% of the reference sequence)[3,4]. ARG-

126   ANNOT[7], Resfinder[9] and DeepARG[10] were able to predict 54, 50 and 2,139 ARDs, respectively, while

127   Resfams[11] predicted a very high number of ARDs (44,105). The HMM-based search for class B1 beta-

128   lactamases published by Berglund *et al.*[22] also yielded a high number of hits (n=3,490) in the 3.9 million

129   protein catalogue (Figure 1C, Supplementary Figure 5). Further analysis on a catalogue of dummy,

130   synthetic 3.9 million proteins indeed showed that Resfams, DeepARG and the Berglund *et al.* HMM-

131   based search lacked specificity (see Supplementary Information). The mean identity shared between

132   predicted (n=6,095) and reference ARDs was 29.8%; it was significantly higher than candidates not

133   predicted as ARDs (mean 23.0%, Wilcoxon unpaired test p=2e-16, Figure 1D). Indeed, most of the

134   pdARDs were distantly related to reference ARDs (Supplementary Figure 6 and 7). Besides, PCM failed

135   to predict 16 ARDs which shared at least 40% identity with a reference ARD (Supplementary Table 2).

136   The 6,095 pdARDs and their structures are available at http://mgps.eu/Mustard.

137

138   **Taxonomic distribution of ARDs**

139   A host bacterial phylum could be assigned to 72.3% (4405/6095) pdARDs. The majority was identified

140   as from the dominant human intestinal phyla Firmicutes (2962/4405, 72.3%) and Bacteroidetes

141   (858/4405, 19.5%) (Supplementary Figure 8) with only 5.8% (225/4405) of pdARDs coming from

142   Proteobacteria. An additional seven pdARDs were predicted to be harboured by Archaea

143   (*Methanobrevibacter* and *Methanoculleus* genera), putatively conferring resistance to macrolides,

144   tetracyclines, aminoglycosides, sulphonamides and glycopeptides (Supplementary Table 1). We also

145   predicted ARDs in genera of medical interest where no ARDs had been identified such as *Akkermansia*[23]

146   (10 pdARDs) and *Faecalibacterium*[24] (44 pdARDs). Only 23 out of 6,095 (0.4%) had been previously

147   identified in families and genera that include human pathogens (Enterobacteriaceae, *Campylobacter*,

148   *Enterococcus*, *Streptococcus* and *Acinetobacter*). The distribution of the families of pdARDs differed

149   according to the phyla (Supplementary Figure 9): Firmicutes and Proteobacteria were enriched with

150   aminoglycosides-modifying enzymes (AMEs, spanning APH, ANT, and AACs) whereas Bacteroidetes

151   were enriched in Sul and class A beta-lactamases. Interestingly, the tigecycline-degrading

152   monooxygenase Tet(X) was frequently found in Bacteroidetes and Proteobacteria, the two phyla

153   between which transfer of the *tet(X)* gene has been reported[14,25]. In order to support these assignments,

6

we sequenced the metagenome of four human faecal samples before and after an overnight culturing using conditions that favoured the growth of oxygen-tolerant bacteria such as Enterobacteriaceae and enterococci (see methods). The results showed an enrichment of Proteobacteria (over Firmicutes and Bacteroidetes), and a commensurate increase of class C beta-lactamases, Fos and Tet(X), along with Van ligases (Supplementary Figure 10).

Location of the pdARDs and association with mobile genetic elements

We investigated the potential for mobility of the pdARDs at different levels. First, we took advantage of the identification of gene clusters based on co-abundance and co-occurrences of genes among the 396 faecal metagenomes used to build the 3.9 million MetaHIT gene catalogue[19]. A total of 7,381 gene clusters referred to as metagenomic units (MGUs) were identified. Among MGUs, metagenomic species (MGS) are defined as MGUs with ≥700 genes, which are considered to be representative of partial or complete bacterial genomes[19]. MGUs of <700 genes include MGEs such as plasmids, phages, transposable elements, and incomplete chromosomal sequences. The 7,381 MGUs from the 3.9 million gene catalogue of intestinal microbiota gene were queried with the pdARDs. A total of 3,651 (59.9%) pdARDs could be mapped onto an MGU. The distribution of pdARDs as a function of MGU size is shown in Figure 2A. Most (95.6%, 3,489/3,651) pdARDs mapped onto MGS and the relative abundance of pdARDs correlated strongly with the abundance of their respective MGS (Supplementary Information), supporting their location on the same bacterial host across the 396 individuals. We also searched for pdARDs in metagenomic species pan-genomes (MSPs)[26] obtained from the 9.9 million intestinal gene catalogue[27]. Similar to MGS, MSPs are clusters of genes that are co-abundant in a set of sample. In MSPs, genes that are constantly found are referred as "core" while inconsistently found genes are referred to as "accessory". Besides, "shared core" genes are assumed to be conserved genes shared between phylogroups[26]. We found 4,912 pdARDs located on MSPs, with the majority being assigned to the core pangenome (4,099/4,912, 83.4%) or shared between core-pangenomes (389/4,912, 7.9%). This was different with MGE-associated genes[27] with most being not found in MSPs (Figure 2B).

Then, we investigated whether genes associated with gene mobility (transposases, conjugative elements and integrons) were present on the same contig than the pdARDs. We found that 7.9% (484/6,095) of pdARDs were co-located with homologs of MGE-associated genes. For pdARDs not

183   found in MGS or in MSPs (n=974), 876 (89.9%) had no detectable MGE-associated genes in their

184   vicinity.

185   Finally, we searched for pdARDs homologs (BLASTN >97% identity over >90% of the query length) in

186   the Genbank database (2018 July 11). Only 538 pdARDs homologs were identified, with 49 being

187   located on a plasmid and/or a phage (Supplementary Table 3). Among the 489 remaining pdARDs, 82

188   (16.8%) were found in multiple species, mainly (60/82, 73.2%) from the same genus (Supplementary

189   Table 4).

190   The phyla Bacteroidetes, Firmicutes and Tenericutes had the higher proportions of ARDs co-locating

191   with MGEs (Figure 2C). No ARD family was found to be enriched in MGE, with the exception of the

192   Tet(X) family in which 3 out of 9 (33.3%) predictions (2 from *Bacteroides fragilis* and 1 from *E. coli*) were

193   associated with transposases (Figure 2D).

194

195   **Distribution of pdARDs in human hosts' microbiota**

196   In the MetaHIT cohort (663 subjects), we found that subjects carried pdARDs with a median relative

197   abundance of 0.22% (range 0.14%-0.38%), with pdARDs from the Tet(M) family being the most

198   abundant (0.07%) and those from class B3 beta-lactamases the least (median: 0.004%). The average

199   number of unique pdARDs genes detected per metagenome was 1,377 (range 258-2,367). Most

200   pdARDs were shared across multiple subjects, 987/6,095 (16.2%) were found in at least 50% of

201   individuals, and only 106/6,095 (1.7%) occurred uniquely in a single individual. All ARD families, with

202   the exception of RNA methylases and AAC(2') families, were found in more than 80% of individuals.

203   Then, we assessed whether subjects with no recent exposure to antibiotics could cluster according to

204   their intestinal resistome. Based on the pdARDs family patterns, six clusters (that we named

205   "resistotypes" by analogy with the enterotypes[28]) were detected using Dirichlet multinomial mixture

206   models (Supplementary Figure 11). The four most frequent resistotypes each represented around 20%

207   of the cohort (the fifth and the sixth representing 8.7% and 7.5%, respectively). The three first

208   resistotypes were characterized by a high abundance of Van ligases (Supplementary Figure 12).

209   Resistotype 1 was enriched in ANT, while resistotype 3 was driven by Tet(M) and class C beta-

210   lactamases. Resistotype 4 was enriched with Tet(X) and class A beta-lactamases and resistotype 6 in

211   class B1 beta-lactamases and Sul. We observed that resistotypes, as determined by PCM, were highly

212   connected to the composition of the microbiota, and that this effect was more pronounced than

213 resistotypes determined from the results of BLASTP and Resfams (Figure 3A). The resistotypes of the

214 MetaHIT cohort were found to be associated with enterotypes (chi square test, p=5e-4), Figure 3B-D,

215 Supplementary Figure 13). Resistotypes 1 and 3 had higher gene richness and were associated with

216 the Clostridiales–driven enterotype. Resistotype 4 was more prevalent in enterotypes driven by

217 Bacteroides (known to harbour Tet(X) and class A beta-lactamases) while resistotype 6 was very

218 specific to the *Prevotella* enterotype (Figure 3C-D). The relative abundance of pdARDs was observed

219 to be positively correlated to the gene richness (Figure 4A, Spearman's rank correlation test Rho=0.31,

220 p=5e-16). Conversely, we did not find any link between resistotypes and body mass index, age or

221 gender.

222

223 **Dynamics of the pdARDs under various exposures to antibiotics**

224 We investigated the abundances of pdARDs in subjects under various exposures to antibiotics and

225 healthcare environments. Three types of exposure were considered (see methods for details):

226 hospitalization in a French hospital without receiving antibiotics, n=15, chronic exposure (Spanish cystic

227 fibrosis patients frequently exposed to antibiotics, n=30) and short high-dose exposure through selective

228 digestive decontamination [SDD; oral colistin, tobramycin, antifungal amphotericin and parenteral

229 cefotaxime[29]] at admission in intensive care units in Netherlands, n=10). We again confirmed a positive

230 correlation between relative abundance of pdARDs and gene richness among patients unexposed to

231 antibiotics (Figure 4B, Spearman's rank correlation test Rho=0.37, p=0.01, see methods). However,

232 when all the samples were considered, including those with antibiotic exposure, this relationship was no

233 longer present (Figure 4C). Instead, the relative abundance of pdARDs was found to be higher in

234 subjects with a chronic exposure than in subjects with no recent exposure (Figure 4D, Wilcoxon unpaired

235 test, p=1e-10), and gene richness was lower (Figure 4E, Wilcoxon unpaired test, p=0.006) In particular,

236 subjects with chronic exposure carried more class B1-B2 beta-lactamases, AAC(6'), ANT, APH, Erm,

237 and DfrA with lower abundance of Sul (Supplementary Figure 14). At the phylum level, we observed a

238 decrease of Bacteroidetes and Verrucomicrobia and an increase of Firmicutes and Actinobacteria in

239 patients chronically exposed to antibiotics (Supplementary Figure 15). A total of 74 MGS were found to

240 be differentially abundant among subjects with or without chronic exposure to antibiotics (Supplementary

241 Table 5).

242   This was different with subjects before and after SDD. A drastic loss of gene richness was measured

243   for this group (Figure 4E): from a mean of 295,919 genes to 95,286 (67.8 % reduction, Wilcoxon paired

244   test, p=0.006). Meanwhile, the relative abundance of pdARDs did not change significantly (Figure 4D,

245   p=0.4). At the ARD family level, we observed that some families decreased significantly: class C beta-

246   lactamases (commonly found in Enterobacteriaceae and Pseudomonadaceae which are specifically

247   targeted by SDD), Fos, Tet(X), APH and ANT (Supplementary Figure 16). We then analysed the MGS

248   at the phylum level and found that Proteobacteria, Actinobacteria, Firmicutes and Fusobacteria

249   decreased significantly after SDD (Supplementary Figure 17). A total of 358 MGS were found in this

250   cohort and, despite the small number of subjects (n=10), we found 133 MGS for which a significant

251   variation was observed (Supplementary Table 6). We tested whether a high abundance of pdARDs

252   could be protective against the antibiotics used in SDD, but found no association: the relative abundance

253   of pdARDs before SDD was not linked to the gene richness after SDD. Hospitalization without antibiotic

254   therapy, that is, potential exposure to antibiotic-resistant nosocomial pathogens without selective

255   pressure, did not affect the gene richness nor the relative abundance of pdARDs (Figure 4D and 4E).

256

257   **Discussion**

258   The results of this study support the concept that the majority of ARDs from the intestinal microbiota are

259   hosted by commensal bacteria, and that their transfer between species (including to opportunistic

260   pathogen) is rare[30]. We provide several findings to support this assumption: 1) we used a 3D structure-

261   based method to assess the diversity of ARDs in the intestinal microbiota and confirmed that ARDs

262   predicted by PCM in the intestinal microbiota were distantly related to known ARDs, 2) the sensitivity

263   and the specificity of the method was validated by gene synthesis of a subset of predictions and by

264   benchmarking against various datasets (functional metagenomic of the soil microbiota, genomes and

265   random protein catalogues), 3) the majority of pdARDs could be found in clusters of co-abundant genes

266   (MGS and MSPs) in large cohorts of samples, while only a minority was found on plasmids, phages or

267   in the vicinity of MGE-associated genes, 4) we could stratify subjects into 'resistotypes' that were

268   connected to enterotypes, and 5) gene richness, otherwise associated with a healthy status[31], was

269   positively correlated to the abundance of ARDs in subjects not exposed to antibiotics.

270   Our results challenge the paradigm that ARDs of the intestinal microbiota are a threat to public health.

271   As was previously demonstrated for environmental samples[18,32], ARDs tend to cluster according to the

272 underlying microbial ecology of the ecosystem, suggesting that the vast majority of ARDs are fixed in

273 their microbial hosts and are not, or very rarely, transferred. Our results show that the dominant intestinal

274 microbiota are not a major conduit through which opportunistic pathogens can acquire ARDs.

275 Nevertheless, we acknowledge that such transfer events have been reported[14,15] and that

276 consequences for public health can be important, as in the case of the *vanB* vancomycin resistance

277 operon that is shared by *Clostridium* spp. and enterococci[15]. Understanding the mechanisms that can

278 lead to the mobilisation of ARDs in the intestinal microbiota, as well as a broader census of

279 environmental reservoirs of ARDs (e.g. sewage, livestock, the subdominant human intestinal

280 microbiota) will continue to be an important area for future research.

281 We found that subjects cluster according to the composition of their resistome into six groups that we

282 named "resistotypes" (as a reference to the previously described enterotypes[28]). These resistotypes

283 were indeed connected to the enterotypes. Description of this underlying structure is interesting as one

284 might hypothesize that a particular resistotype, or microbiota enriched with ARDs, might be affected to

285 different degrees by antibiotic therapy. This has previously been observed for beta-lactamase-producing

286 *Bacteroides* which can protect the microbiome against exposure to β-lactams[33]. In patients undergoing

287 faecal microbiota transplantation, follow-up antibiotic therapy may be adjusted to favour engraftment of

288 the donor microbiota[34]. Identifying donors with a resilient microbiota, due to a protective resistotype,

289 could open perspectives for the optimisation of the clinical implementation of faecal microbiota

290 transplants.

291 Contrary to initial expectations, some pdARD families decreased in their abundance under antibiotic

292 exposure, especially when patients were exposed to a combination of antibiotics (such as SDD). In order

293 to resist to a combination of antibiotics, bacteria would need to be intrinsically resistant or to acquire an

294 adequate combination of ARDs. The dynamics of ARDs under antibiotic exposure depend on various

295 parameters: spectrum of the ARD (the level of resistance towards the antibiotic provided by the ARD),

296 the expression level of the ARD, and the presence of other resistance mechanisms (intrinsic or

297 acquired). The large number of possible combinations of these factors can explain that in some

298 situations, a bacterium can be inhibited by antibiotics despite the presence of a putatively compatible

299 ARD. Alternatively, we cannot exclude that changes in pdARDs families could also be explained by

300 simple taxonomic shifts that are not connected to the antibiotics studied.

301   The limitations of current techniques and of this study leave a number of important questions unresolved.

302   As mentioned earlier, metagenomic sequencing provides information for the dominant fraction of

303   intestinal bacteria, and so ARDs present in subdominant bacteria remain unobserved. Indeed, several

304   ARDs found in opportunistic pathogens among the Enterobacteriaceae (e.g. *Escherichia*

305   *coli* and *Klebsiella pneumoniae*) originate from other species in the same Proteobacteria phylum[35]. A

306   recent study indeed cultured many Proteobacteria species that were not detected in metagenomic

307   sequencing[36]. We cannot rule out that the subdominant bacteria, which were not probed by

308   metagenomic sequencing, could be an additional reservoir of ARDs. In terms of the clinical samples

309   analysed, we cannot exclude that the differences between patients and controls may be resulting from

310   confounding factors other than the antibiotic exposure.

311   The method we used to identify distantly related proteins is based on homology modelling and takes

312   advantage of the observation that proteins sharing the same function have more similar structures than

313   amino acid sequences[37]. Indeed, PCM could identify functional ARDs with amino acid identity below

314   20% to known ARDs. Notably, PCM can only be used to predict the function of genes that are

315   homologous to known ARDs, and therefore the identification of different classes of ARDs with no

316   homology to known ARDs will still require functional screening. Besides, while PCM was validated in

317   this study, it remains a prediction tool. While similar structures are usually indicative of similar function,

318   this is not always the case and PCM can yield false positives results (as observed in the functional

319   validation of synthesized pdARDs). Due to the scope of our study, gene synthesis validation was not

320   performed for all ARD families, leaving open the possibility that not all pdARDs identified here truly have

321   a role in antibiotic resistance.

322   In summary, we developed a method, PCM, which could unveil the diversity of ARDs in the intestinal

323   microbiota. Employing this tool, we gathered evidence that the vast majority of the ARDs we predicted

324   showed no sign of mobility and that their abundance was correlated to gene richness. Together with the

325   protective trait of some intestinal bacteria against antibiotics[33], our results suggest that the ARDs from

326   the intestinal microbiota might be considered as our "resilience allies"[38] assuring the preservation of the

327   healthy commensal microbiota under antibiotic exposure.

328

329   **Acknowledgements**

348

349   **Conflicts of interest**

350   None.

351

352   **Authors' contribution**

353   ER, AG, JT performed the analysis. ER, AG, JT, WvS, AdB and SPK wrote the manuscript. ASA and

354   NM handled the data management. TC, SHA, IC and JLM performed the gene synthesis experiments.

355   JLM, TMC, VFL, FB, AdB, JD, SPK, FH and SDE discussed the protocol and results. LM, TG, VdL, NA,

356   BF, IW, AA, WvS, MR, XZ and RJL recruited the patients and collected the samples. HB, VL, AL and

357   FL handled the wet lab experiments. NP, PL and JMB managed the informatics and the calculation

358   clusters. KW and NP designed the website (http://mgps.eu/Mustard/).

359

360

Methods

Constitution of the databases of antibiotic resistance determinants

We define as an ARD as in Martinez *et al*[39]: a protein encoded by a gene that confers resistance to antibiotics when it is present or increases susceptibility to antibiotics when it is absent. This definition excluded housekeeping genes in which mutations can confer resistance to some antibiotics (such as topoisomerases in which mutations can lead to fluoroquinolone resistance) and genes involved in the regulation of antibiotic resistance genes. Also, we excluded efflux pumps such as TetA or QepA as very few or no PDB are available, presumably due to the difficulty to crystallize transmembrane proteins. Amino acid sequences of functionally characterized ARDs from the major antibiotic families used in human medicine (beta-lactams, aminoglycosides, tetracyclines, trimethoprim, sulfonamides, macrolides-lincosamides-synergistines, fluoroquinolones, fosfomycin and glycopeptides)[20,40] were obtained from the following antibiotic resistance databases: Resfinder[9], ARG-ANNOT[7], the Lahey Clinic (http://www.lahey.org/studies/), RED-DB (http://www.fibim.unisi.it/REDDB/), Marilyn Roberts's website for macrolides and tetracycline resistance genes (http://faculty.washington.edu/marilynr/) and from functional metagenomics studies[5,6,41]. When ARDs were provided as nucleic acids sequences, they were translated into proteins with Prodigal[42]. Non-redundancy of the reference ARDs was assessed with CD-HIT v4.5.7[43] (100% identity). The final database was manually curated in order to remove incomplete sequences and ARDs from families not considered in this work. The cluster of orthologous genes (COG) of each member of the reference dataset was assigned from the v3 eggNOG database[44]. In total, we collected 1,651 non-redundant amino acid sequences spanning 20 ARDs families: Class A beta-lactamases (Blaa), class B1-B2 beta-lactamases (Blab1), class B3 beta-lactamases (Blab3), class C beta-lactamases (Blac), class D beta-lactamases (Blad), aminoglycoside acetyltransferases (AAC) AAC(2'), AAC(3)-I, AAC(3)-II, and AAC(6'), aminoglycoside nucleotidyltransferases (ANT), aminoglycoside phosphotransferases (APH), 16S rRNA methylases, Tet(M), Tet(X), type A dihydrofolate reductases (DfrA), dihydropteroate synthases (Sul), erythromycin ribosome methylases (Erm), quinolone resistance proteins (Qnr), fosfomycin resistance proteins (Fos), and D-Ala – D-Lac/Ser ligases (Van) (Table 1). The recently described plasmid-mediated colistin resistance *mcr-1* gene[45] could not be included because of the lack of a reliable PDB template obtained by X-ray diffraction at the time of the study.

Interrogation of the catalogue for ARDs

We used a 3,871,657 million proteins catalogue previously published[19]. This catalogue was built from the metagenomic sequencing of the faeces of 396 subjects from Denmark and Spain. In brief, the 3.9 million gene catalogue results from a non-redundancy filtering at 95% nucleic acid identity and 90% coverage: predicted genes from all samples (45.4 million in total) were clustered using BLAT by single linkage. Any two genes with greater than 95% identity and covering more than 90% of the shorter gene were clustered together. The contigs were originally built using SOAPdenovo (from the MOCAT pipeline[52]). We selected this catalogue over the more recent 10 million gene catalogue that was published during the course of this study[27] because metagenomic units (MGUs, including the metagenomic species [MGS]) had been determined only for the 3.9 million gene catalogue. The genes of the catalogue were translated into proteins using Prodigal[42] using the –p meta option. For each ARD family, we searched for ARDs using the three following methods: (*i*) we built a hidden Markov model file for each ARD family and searched the catalogue with Hmmsearch (v3. 1)[46], (*ii*) we performed a Smith-Waterman alignment with a heuristic seed detection (BLASTP v. 2. 2. 28+)[21] and (*iii*) a rigorous Smith-Waterman search (SSearch v. 36. 3. 6)[47] with an E-value threshold of 1E-5. Only the hits with a size ranging from 75% and 125% of the mean amino acid size of the ARD family were further considered. All candidates were assigned a COG/NOG from eggNOG v3[44]. When candidates were found in different ARD families (e.g. a candidate could be a hit in class B1-B2 and class B3 beta-lactamases), the candidate was assigned to the family for which it had the highest amino acid identity with the reference.

Negative references

For each ARD family, COGs/NOGs were attributed to reference ARDs. In parallel, the COGs/NOGs were attributed to the hits obtained during the initial steps of PCM (i.e. the hits obtained by the BLASTP/SSearch and Hmmer search). In the list of candidates from a given ARD family, the COGs/NOGs that were not found in the COGs/NOGs attributed to reference ARDs were assumed to be potential COGs/NOGs from false positives hits (Supplementary Figure 2) as it reproduced the errors of functional assignment likely to be generated in sequence-only annotations. The amino acid sequences of the representative proteins from those COGs/NOG groups were obtained from the eggNOG v3 database, and were added to the negative reference dataset. A manual curation step was performed in order to ensure that no references were included in the negative references.

421

Selection of structural templates

The list of protein structures that could be used as structural templates was downloaded (June 2014,

and November 2014) from the PDB library (Protein DataBank[48], http://www.rcsb.org/). Using the

reference dataset and the negative references described above, Hmmer[46], BLASTP[21] and SSearch[47]

were performed on the PDB database with default settings and E-values of 1E-5. Results were merged

into a non-redundant PDB list. Both lists (references and negative templates) were manually curated to

ensure that no references were represented in the negative templates dataset, and vice versa. If more

than one PDB shared the same UniProt number (i.e. if the structure of a protein has been determined

on multiple occasions), we filtered the PDB files in order in include a unique structure per UniProt number

using the following positive criteria: absence of ligand, completeness of the protein and high resolution.

Pairwise comparative modelling

The concept of pairwise comparative modelling (PCM) is shown in Supplementary Figure 1-3 and the

framework is available at https://github.com/aghozlane/pcm. The concept of

leveraging the protein structure in complement to its amino acid sequence was motivated by the fact

that proteins sharing common functions would be more conserved in the active site which cannot be

observed by the analysis of protein sequence alignments [37]. Each candidate was subjected to homology

modelling with reference templates and negative templates, generating two 3D structures for each

candidate (Fig 1A). The main idea is that if a sequence is truly functionally related to the reference fold,

its model must be significantly different from the ones obtained with the negative structural template.

Homology modelling was performed by PCM in six main steps (example in Supplementary Figure 3):

1.   Three structural templates were identified by BLASTP (among the lists produced as described

above) that shared the highest amino acid identity with the candidate protein.

2.   A multiple sequence alignment was performed between the candidate and the three templates

sequences using Clustalo[49].

3.   A prediction of the secondary structure was performed using psipred (v3.5)[50]. The residues

predicted to fold in helix or in beta-sheet conformation with a level of confidence higher or equal to 7

were considered to constrain the model.

450   4.   A comparative modelling was performed with the MODELLER programming interface[51].

451   MODELLER automatically calculates a model by satisfaction of spatial restraints such as atomic

452   distance and dihedral angles in the target sequence, extracted from its alignment with the template

453   structures. Stereo-chemical restraints for residues are obtained from the CHARMM-22 molecular force

454   field and statistical preferences obtained from a representative set of known protein structures.

455   5.   The best model out of a hundred produced by MODELLER (based on the Dope score) was

456   considered for structure assessment analysis using ProQ[52] and Prosa-web[53]. The Dope score

457   (Modeller), z-score (Prosa), MaxSub and LG score (ProQ) are statistical potential variables used to

458   predict the model quality. Both ProQ and Prosa-web are trained on the PDB to determine real protein

459   configuration and they estimate the energetic favourability of the conformation of each residue in the

460   model.

461   6.   The best model was aligned with the reference set of structures using TM-align[17] and

462   MAMMOTH[54]. The RMSD (TM-align), z-score (MAMMOTH), TM-score (MAMMOTH, TM-align)

463   estimates the degree of superposition of the residue between two structures.

464   The differences (delta) between the scores determined from each modelling path (with the reference set

465   or the negative set) were calculated and used for the PCM machine learning program (see below).

466   For one given candidate, the PCM whole process took an average of 8 CPU-hours (30 minutes on 16

467   CPUs).

468

469   **Taxonomic assignation**

470   pdARDs were taxonomically assigned by combining the results obtained from BLASTN against the NCBI

471   Genomes database (minimal 70% identity and 80% coverage), a BLASTN against the IMOMI in-house

472   database (minimal 85% identity and 90% coverage) and the taxonomy of the metagenomic unit

473   whenever applicable. The lowest taxonomic rank from the results of the three methods was assigned to

474   the pdARD.

475

476   **Statistical analysis**

477   To discriminate reference proteins from negative references, we used model quality predictors and

478   alignment scores (inferred from the semi-automatic pipeline described above) and developed a custom

479   pipeline in R (R Core Team, 2013, http://www.R-project.org) to perform the classification. The LASSO

480   penalized logistic regression[55] implemented in LIBLINEAR[56] was used to compute the classifier. Ten-

481   fold stratified cross validation (re-sampled 100 times to obtain more stable accuracy estimates) was

482   used to partition the data into a training and test sets. The LASSO hyper-parameter was optimized for

483   each model in a nested 5-fold cross-validation on the training dataset using the area under curve (AUC)

484   as the model selection criterion. From the 100 times re-sampled ten-fold cross validation, receiver

485   operating characteristic (ROC) analysis was used to evaluate model performance using the median

486   AUC. Coefficients extracted for each modelling or alignment score were also evaluated for their stability

487   throughout the computed models. The PCM score was the ratio (expressed as a percentage) between

488   the numbers of time a candidate was classified as a reference and the number of bootstraps. Predicted

489   ARDs were candidates with a PCM score ≥50% and a TM score given by TM-align ≥0.5[17]. To control

490   how structural modelling brought additional information compared to amino acid sequence alignment

491   only, we built a logistic regression model based on T-coffee alignment score (R glm, ten-fold

492   stratification, re-sampled 100 times). We then compared the two classifiers models used for PCM and

493   for T-coffee alignment based on the reference set (see Supplementary Information).

494

Validation of the method with a functional metagenomic dataset

496   The performance of PCM was assessed by analysing the data in Forsberg *et al.*, where the ARD content

497   of different North American soils was analysed using functional metagenomics[18]. The screening of the

498   clones was performed on aztreonam, chloramphenicol, ciprofloxacin, colistin, cefepime, cefotaxime,

499   cefoxitin, D-cycloserine, ceftazidime, gentamicin, meropenem, penicillin, piperacillin, piperacillin-

500   tazobactam, tetracycline, tigecycline, trimethoprim and trimethoprim-sulfamethoxazole (cotrimoxazole).

501   Here, we collected the nucleotide sequences of the inserts deposited on Genbank (KJ691878–

502   KJ696532). The sequence translation of the open reading frames was performed by Prodigal (using

503   default parameters)[42]. A total of 4,654 insert sequences were collected, in which 12,904 amino acid

504   sequences were predicted. We then searched for ARDs belonging to the relevant ARD families

505   according to the antibiotics used for the screening of the clones: beta-lactamases (all classes), APH,

506   ANT, AAC(2’), AAC(3)-I, AAC(3)-II, AAC(6’), RNA methylases, Tet(M), Tet(X), Qnr, Sul and DfrA, using

507   the Supplementary Table 2 of the Forsberg *et al.* paper. Inserts with no putative ARDs (according to the

508   annotation of the gene) were removed (n=269). Inserts selected on cycloserine (n=868) and

509   chloramphenicol (n=129) were not considered here because they were not included in the 20 ARD

families in this work. Fourteen inserts which contained more than one putative ARD that could be identified to confer resistance to the antibiotic used for the screening (e. g.; two beta-lactamases) were not considered in this analysis. An additional 1,658 inserts containing no putative ARDs or a putative ARDs that did not confer resistance to the antibiotic used for selection were discarded and so were 294 inserts containing efflux pumps, as these were not considered in this study. The resulting validation set contained 1,423 inserts (with resistance genes) for a total of 3,778 genes. To compare the outcome of PCM with other tools, the results for class B1-B2 and B3 beta-lactamases generated by PCM were merged into one class B beta-lactamases group as other tools do not separately consider the different class B beta-lactamases.

In total, 1,390 unique hits were found during the initial screen of PCM, of which 1,374 were predicted as ARDs (Supplementary Table 7). Among the 33 ARDs not included for PCM, 12 were not considered because they were undersized and 10 because they were oversized. No hits for AAC(2'), ANT, Qnr or Sul were found. The mean identity shared with reference ARDs was 37.6% (range 18.8-94.5). Overall, the sensitivity was 96.6%, with no false negative. In comparison, only 8 ARDs would have been identified by a conventional method (combination of Hmmsearch, BLASTP and SSearch with both a minimal identity with a reference ARD and coverage over or equal to 80%). Conversely, Resfams[11] that was specifically designed to identify ARDs from functional metagenomic datasets showed a similar sensitivity to PCM with the identification of 1,346 ARDs out of 1,423 (94.6% sensitivity).

Validation of the method for incomplete genes

The 3.9 million gene catalogue harbours 41.4% of genes that are predicted to be incomplete either on the 5', the 3' or both extremities[19]. As the size parameter is crucial for homology modelling, we tested to what extent the prediction of incomplete ARDs by PCM could remain valid. We selected 12 reference class A beta-lactamases (BlaZ, CblA-1, CepA-29, CfxA2, CfxA6, CTX-M-8, KPC-10, OXY-1, PER-1, SHV-100, TEM-101 and VEB-1) and we then iteratively removed 5% of the amino acid sequence at both edges in order to obtain 16 bi-directionally trimmed candidates (from 100% to 25%) per reference ARD. Candidate genes were chosen to span the diversity of known beta-lactamases, but the main representative beta-lactamase of the subfamily (e.g. TEM-1 for TEM beta-lactamase) was not necessarily chosen. Note that SHV-100 has a slightly longer sequence (13 amino acid duplication) than other SHV. A total of 192 PCM experiments were performed: we observed that the 12 references were

correctly predicted as ARDs when at least 40% of the protein remained (i.e. 30% trim from each extremity, Supplementary Figure 4). Thus, we are confident that with the 75% size threshold used in this study (a maximum of 25% removed from one edge), no misclassification due to an incomplete gene would be expected.

Gene synthesis

We selected 71 pdARDs from 12 ARD families: 14 from class A beta-lactamases, 8 from class B1-B2 beta-lactamases, 7 from class B3 beta-lactamases, 4 from class C beta-lactamases, 2 from class D beta-lactamases, 2 AAC(3)-I, 5 AAC(3)-II, 8 AAC(6'), 3 ANT, 4 APH, 13 Tet(M) and 1 Tet(X)) for gene synthesis and sub-cloning into *Escherichia coli* to test the decrease of susceptibility to antibiotics. For beta-lactamases, a chromogenic test (nitrocefin) was used to detect function. Minimal inhibitory concentrations (MIC) were determined by E-Test strips (bioMérieux, Marcy-l'Etoile, France) in duplicate. A pdARD was considered to have an activity against an antibiotic (tobramycin for AAC(3)-I, AAC(3)-II, AAC(6') and ANT; kanamycin for APH and tetracycline for Tet(M)) when the MIC of the clone was above the MIC of a clone harbouring the plasmid without a synthesized gene or when the colour of the broth containing nitrocefin turned red, in the case of beta-lactamases. We used the plasmid vector pET-22b+ (embedding a beta-lactamase – encoding gene) for pdARDs hypothesized to confer resistance to aminoglycosides and the pET-26b (embedding a gene conferring resistance to kanamycin) for the other pdARDs. The selection of the pdARDs for synthesis was performed as follows:

- References (n=12): pdARDs which shared a high identity with known ARDs (≥95% amino acid identity and ≥80% coverage with a reference ARD).

- Good predictions (n=41): pdARDs with the highest degree of confidence for the prediction (PCM score >99%, Tm score TmAlign>0.9 and <70% amino acid identity with a reference ARD.

- Fair predictions (n=18): pdARDs with the lowest degree of confidence for the prediction (PCM score <80%, Tm score TmAlign<0.8 and <70% amino acid identity with a reference ARD).

Signatures of mobile genetic elements nearby the predictions of ARDs

We searched for mobile genetic elements (MGE) - associated proteins encoded by genes located in the same contigs as pdARDs. The 3.9 million gene catalogue results from a non-redundancy filtering at 95% for the genes[19], but in order to identify the contigs on which pdARDs were identified, we needed to return

570   to the redundant catalogue (*i.e.* the non-dereplicated catalogue of genes) and identified homologs

571   sharing 95% nucleic acid identity with the pdARDs. By doing so, we could identify contigs (n=16,955)

572   carrying at least one pdARD. The mean size of the contigs was 19,711 bp (min 500, max 461,981,

573   median 8,513). In total, the 16,955 contigs contained a total of 908,888 genes after the subtraction of

574   pdARDs. The 908,888 genes were then translated into proteins with Prodigal[42] and queried for IS

575   elements using BLASTP (query size threshold 150 amino acids, e-value 1E-30, identity threshold 40%)

576   against the ISfinder database[57]. Conjugative elements were queried among the same gene set

577   (n=908,888) with Conjscan[58], using the default parameters and the filters recommended by the authors

578   (best e-value<0.001 and sequence coverage of at least 50%). Most proteins belonging to the type IV

579   secretion systems (T4SS), which are involved in conjugation, are ubiquitous in that they have numerous

580   homologs. Hence, when searching for conjugation proteins in a 3.9 million protein catalogue, there

581   would be a high risk of false positives. Accordingly, the colocation of hits was deemed crucial. A

582   conjugative T4SS is made from:

583   • a protease (VirB4)

584   • a second coupling protein protease (t4cp)

585   • a relaxase (MOB)

586   • a proteic complex (MPF) composed of at least 10 proteins

587   In order to identify a T4SS on a contig, we required presence of at least 1 virB4 hit, a t4cp1 or t4cp2 hit,

588   a MOB hit and a certain number of MPF hits. All hits must co-localize. A MOB element alone can mobilize

589   a neighboring gene (such as an ARD-encoding gene) via other T4SSs. However, in our dataset the

590   short length of contigs led us to adapt those parameters (following the recommendations of the

591   developers of the Conjscan software). Besides the MOB element, we considered that the presence of 2

592   hits from the same family (e.g. T_virB6 and T_virB8, or B_traF and B_traH) or virB4+any hit from another

593   family on the same contig as a pdARD was a strong indication of the presence of mobility associated

594   elements. Integrons were identified using IntegronFinder[66] on the 16,955 contigs using default

595   parameters.

596   We also searched for pdARDs in metagenomic species pan-genomes (MSPs)[26] obtained from the 9.9

597   million intestinal gene catalogue[27] using BLASTN with a 95% identity threshold over 90% of the query.

598   We also searched for homologs of pdARDs in Genbank with 97% identity threshold over 90% of the

599   query. We found 820 out of 6095 pdARDs (13.5%) which aligned against 139,413 Genbank entries. We

600   filtered hits corresponding to a virus, a plasmid or a vague taxonomic affiliation by considering the

601   following terms: "uncultured bacterium", "artificial", unidentified", "uncultured organism", "environmental

602   samples" and "metagenome".

603   Distribution of the pdARDs in the MetaHIT cohort (n=663 subjects)

604   pdARDs profiles were obtained from the abundance matrix of the 3.9 million genes as described in

605   Nielsen *et al*[19]. The "reads per kilobase per million mapped reads" (RPKM) method was used to

606   normalize the mapping counts. After summing the relative abundances of pdARDs genes belonging to

607   the same family, Dirichlet multinomial mixture models were used to find ARDs clusters (*i.e.* resistotypes)

608   using the Dirichlet Multinomial R package. The same method was applied to detect gut microbiota

609   clusters *(i.e.* enterotypes)[59]. The Laplace criterion was used to define optimal number of clusters as

610   described on oral and faecal microbial dataset[60]. By analogy with the term enterotype, we chose to name

611   a cluster of subjects based on their similarity of their faecal relative abundance of pdARDs a

612   "resistotype". The Chi-squared test was used to assess the associations between resistotypes and

613   enterotypes. Rarefaction analysis at one million reads was done to determine the gene richness per

614   samples. RLQ analysis[61] was conducted to assess the associations between the relative abundances

615   of pdARDs*,* their characteristics (family, size of the cluster of associated genes [CAG]) and those of

616   subjects (enterotypes, resistotypes, gender, body mass index [BMI], age). Of note, we excluded the

617   patients suffering from inflammatory bowel disorders from this analysis. Co-inertia analysis was

618   conducted to assess the associations between microbiota beta-diversity and pdARDs profiles.

619   Microbiota composition was assessed using metagenomics species (MGS, see below) relative

620   abundance and beta-diversity by square root Jensen-Shannon Divergence (JSD). A principal coordinate

621   analysis was done on JSD distance matrix and a principal component analysis was done on ARDs

622   profiles. Both analyses were then subjected to co-inertia analysis and Monte-Carlo permutation was

623   done to asses to robustness of shared inertia.

624

625   Constitution of cohorts of patients with various antibiotic exposures

626   We included three cohorts of patients with various exposures to antibiotics:

627   - Hospitalization without antibiotics: a total of 31 patients with no exposure to antibiotics or hospitalisation

628   during the three preceding months and admitted to the medicine ward of the Beaujon University

629 Teaching Hospital (Clichy, France) were included and provided a faecal sample at admission. Among

630 them, 16 also provided a stool sample at discharge. One patient received antibiotics between admission

631 and discharge and was not further considered for the analysis. In total, 15 patients could provide a stool

632 sample soon after admission (T0) and at discharge (T1). The mean time between T0 and T1 samples

633 was 10.7 days. The mean age of patients was 67.8 years old and the gender ratio (M/F) was 1.3. All

634 patients gave informed consent. This work was approved by the French National Institutional Review

635 Board (IRB 00008522) and registered at clinicaltrials.gov (NCT02031588).

636 - Chronic exposure: 30 cystic fibrosis (CF) patients were enrolled at the Cystic Fibrosis Unit of the

637 Ramón y Cajal Hospital in Madrid. One faecal sample was collected at the occasion of a consultation.

638 All subjects for this study were provided a consent form describing the study and providing sufficient

639 information for subjects to make an informed decision about their participation as faecal donors in this

640 study. Cystic fibrosis is a genetic disease that leads to an impairment of the lung function through an

641 uncontrolled production of mucus. The consequence is chronic bacterial colonization, resulting in

642 deleterious reactive fibrosis of the lung. Bacterial load is controlled by chronic exposure to antibiotics

643 (home-therapy, mostly oral and inhaled in our cohort), which has resulted in significant life prolongation,

644 and the near-absence of hospital care. Hence, the CF patients had been exposed to various antibiotics

645 during the five years before the faecal sample was collected:

646 ▪ Beta-lactams (ampicilln, amoxycillin, cloxacillin, piperacillin-tazobactam, cefepime, ceftriaxone,

647 ceftazidime, cefditoren, meropenem): 25/30

648 ▪ Macrolides (azithromycin, clarithromcyin): 17/30

649 ▪ Colistin: 21/30

650 ▪ Fluoroquinolones (ciprofloxacin, levofloxacin, moxifloxacin): 26/30

651 ▪ Cotrimoxazole: 14/30

652 ▪ Glycopeptides (vancomycin): 1/30

653 ▪ Aminoglycosides (amikacin, tobramycin): 12/30

654 ▪ Tetracyclines (doxycycline, minocycline): 2/30

655 ▪ Linezolid: 3/30

656 ▪ Rifampin: 1/30

657 ▪ Fosfomycin: 5/30

658 On average, CF patients had been exposed to 5.9 different antibiotics and had an average of 12.2

659 antibiotic courses during the five years before the sample was taken. The mean age was 36.3 years old

660 and the gender ratio (M/F) was 1.3. This protocol and any amendments were submitted to the Ethics

661 Committee (EC) in agreement with local legal prescriptions, for formal approval of the study conduct.

662 The consent form was obtained before that subject provided any faecal sample for the study and was

663 signed by the subject or legally acceptable surrogate, and the investigator-designated research

664 professional obtaining the consent. According to the National Spanish laws the study did not require the

665 approval of the Ethics Committee. Nonetheless, the Ethics Committee of the Hospital Ramón y Cajal

666 guaranteed that the study was performed done according to the good clinical practices guidelines.

667 - Short high dose exposure: selective digestive decontamination (SDD) consists in administering a

668 mixture of topical and parenteral antibiotics and antifungal agents to a patient at admission in order to

669 eliminate potential bacterial and fungal pathogens. SDD has been showed to significantly reduce

670 mortality in the intensive care unit (ICU)[29] and is now part of standard care for intensive care patients in

671 the Netherlands. To assess the effect of SDD on the intestinal microbiota, we analysed the faecal

672 samples from 13 patients admitted to the ICU of the University Medical Centre of Utrecht (UMCU,

673 Netherlands). The samples were collected at admission (T0, first sample passed after admission) and

674 after SDD (T1). Among the 13 patients for whom a faecal sample could be obtained at T0, 10 could

675 provide a faecal sample at T1. The mean age was 59.9 years old and the gender ratio (M/F) was 0.5.

676 SDD consisted of 4 days of intravenous cefotaxime and topical application of tobramycin, colistin, and

677 amphotericin B. Additionally, a subset of samples (n=4) from this cohort was cultured in a brain-heart

678 infusion broth overnight in ambient atmosphere at 37°C. The protocol for the collection of stool samples

679 was reviewed and approved by the institutional review board of the University Medical Centre of Utrecht

680 (The Netherlands) under number 10/0225. Informed consent for faecal sampling during hospitalization

681 was waived. Written consent was obtained for the collection of faecal samples after hospitalization.

682

683 **Metagenomic sequencing and mapping.**

684 Total faecal DNA was extracted[62,63] and sequenced using SOLiD 5500 wildfire (Life Technologies)

685 resulting in a mean of 68.5 million sequences of 35-base-long single-end reads. High-quality reads were

686 generated with quality score cut-off >20. Reads with a positive match with human, plant, cow or SOLiD

687 adapter sequences were removed.

688   Filtered high-quality reads were mapped to the MetaHIT 3.9 million gene catalogue[19] using the METEOR

689   software[64]. The read alignments were performed in colourspace with Bowtie software (version 1.1.0)[65].

690   Uniquely mapped reads (reads mapping to a single gene from the catalogue) were attributed to the

691   corresponding genes. Shared reads (mapping different genes of the catalogue) were attributed

692   according to the ratio of their unique mapping counts, as following: as a read can map on different genes

693   of the catalogue, the abundance of a gene $G(A_g)$ depends on the abundance of uniquely mapped reads

694   (A_u), *i.e.* reads that map only to the gene G, and on the abundance of N shared reads (A_s) that aligned

695   with M genes in addition to the gene G:

696   $$A_g = A_u + A_s$$

697   where

698   $$A_s = \sum_{i=1}^{N} C_{o_i}$$

699

700   For each shared read, the gain of abundance corresponds to a coefficient C_o that takes in account the

701   total number of uniquely mapped reads on the M genes:

702   $$C_{o_i} = \frac{A_u}{A_u + \sum_{j=1}^{M} A_{u_j}}$$

703

704   For instance, if a gene G is mapped by 10 reads that only map to it (unique reads), but also with 1 read

705   that also align on a gene M that was mapped by 5 unique reads, then:

706   $$A_g = 10 + \frac{10}{10 + 5} \approx 10.7$$

707

708   To decrease technical biases due to different sequencing depth, samples with at least 5 million mapped

709   reads were downsized to 5 million mapped reads (random sampling of 5 million mapped reads without

710   replacement) using R package momr[31]. The abundance of each gene in a sample was then normalized

711   by dividing the number of reads that mapped to the gene (A_g) by the gene nucleotide length and by the

712   total number of reads from the sample. The resulting set of gene abundances, termed a "microbial gene

713   profile", was used to estimate the abundance of metagenomic species (MGS)[19].

714

715   *Gene richness analysis*

716   Microbial gene richness was calculated by counting the number of genes mapped at least once for a

717   given sample. Gene richness was calculated using R package momr for samples where 5 million or

718   more reads had been mapped to the 3.9 million gene catalogue.

719

720   *MetaGenomic Species (MGS)*

721   MGS are co-abundance gene groups with more than 700 genes and can be considered as part of

722   complete bacterial species genomes. 741 MGS were delineated from 396 human gut microbiome

723   samples[19]. In this study, the relative abundance of MGS was determined as the median abundance of

724   90% of the genes composing each cluster, meaning that the 10% genes with the lowest abundance for

725   each MGS were not considered for the calculation of the abundance of the MGS. Typically, these genes

726   correspond to genes with 0 count, to accessory genes (hence their detection is not constant) or to genes

727   that are not detected because of insufficient sequencing depth. The MGS taxonomical annotation was

728   updated by sequence similarity using NCBI BLASTN, when more than 50% of the genes matched the

729   same reference of NCBI database (December 2014 version) at a threshold of 95% of identity and 90%

730   of gene length coverage to get the species annotation[19].

731

732   *Statistical analysis for the distribution of pdARDs and MGS between groups*

733   Statistical analyses for the differential abundances of pdARDs and MGS were performed using the

734   application   SHAMAN[66]   (http://shaman.c3bi.pasteur.fr/).   Data   are   available   at

735   (https://github.com/aghozlane/evotar), with the graphical representations using the abundances from

736   the matrix rarefied at 5M reads. The relationship between richness and the abundance of ARDs was

737   assessed by Spearman correlation test. The statistical threshold for significance was set at a p-value of

738   0.05.

739

740   *Data availability*

741   The 6,095 pdARDs PDB files, nucleotide and amino acid sequences can be downloaded from

742   http://mgps.eu/Mustard/. The 3.9 million gene catalogue and the metagenomic species database are

743   accessible at https://www.cbs.dtu.dk/projects/CAG/. The reads from the clinical samples generated in

744   this study are available under the accession number PRJEB27799 at the European Nucleotide Archive

745   (ENA).

746

749

References

750

751

752   1.   Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing.

753   *Nature* **464,** 59–65 (2010).

754   2.   Ghosh, T. S., Gupta, S. S., Nair, G. B. & Mande, S. S. In silico analysis of antibiotic resistance

755   genes in the gut microflora of individuals from diverse geographies and age-groups. *PloS One* **8,**

756   e83823 (2013).

757   3.   Hu, Y. *et al.* Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human

758   gut microbiota. *Nat. Commun.* **4,** 2151 (2013).

759   4.   Forslund, K. *et al.* Country-specific antibiotic use practices impact the human gut resistome.

760   *Genome Res.* **23,** 1163–1169 (2013).

761   5.   Sommer, M. O. A., Dantas, G. & Church, G. M. Functional characterization of the antibiotic

762   resistance reservoir in the human microflora. *Science* **325,** 1128–1131 (2009).

763   6.   Moore, A. M. *et al.* Pediatric fecal microbiota harbor diverse and novel antibiotic resistance genes.

764   *PloS One* **8,** e78822 (2013).

765   7.   Gupta, S. K. *et al.* ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes

766   in bacterial genomes. *Antimicrob. Agents Chemother.* **58,** 212–220 (2014).

767   8.   McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents

768   Chemother.* **57,** 3348–3357 (2013).

769   9.   Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob.

770   Chemother.* **67,** 2640–2644 (2012).

771   10. Arango-Argoty, G. *et al.* DeepARG: a deep learning approach for predicting antibiotic resistance

772   genes from metagenomic data. *Microbiome* **6,** 23 (2018).

773   11. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance

774   determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9,** 207–216 (2015).

775 12. Wright, G. D. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat. Rev.*

776 *Microbiol.* **5,** 175–186 (2007).

777 13. Salyers, A. A., Gupta, A. & Wang, Y. Human intestinal bacteria as reservoirs for antibiotic

778 resistance genes. *Trends Microbiol.* **12,** 412–416 (2004).

779 14. Ghosh, S., Sadowsky, M. J., Roberts, M. C., Gralnick, J. A. & LaPara, T. M. Sphingobacterium sp.

780 strain PM2-P1-29 harbours a functional tet(X) gene encoding for the degradation of tetracycline. *J.*

781 *Appl. Microbiol.* **106,** 1336–1342 (2009).

782 15. Stinear, T. P., Olden, D. C., Johnson, P. D., Davies, J. K. & Grayson, M. L. Enterococcal vanB

783 resistance locus in anaerobic bacteria in human faeces. *Lancet* **357,** 855–856 (2001).

784 16. Penders, J., Stobberingh, E. E., Savelkoul, P. H. M. & Wolffs, P. F. G. The human microbiome as

785 a reservoir of antimicrobial resistance. *Front. Microbiol.* **4,** (2013).

786 17. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score.

787 *Nucleic Acids Res.* **33,** 2302–2309 (2005).

788 18. Forsberg, K. J. *et al.* Bacterial phylogeny structures soil resistomes across habitats. *Nature* **509,**

789 612–616 (2014).

790 19. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex

791 metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32,** 822–828 (2014).

792 20. Goossens, H., Ferech, M., Vander Stichele, R., Elseviers, M. & ESAC Project Group. Outpatient

793 antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet*

794 **365,** 579–587 (2005).

795 21. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search

796 tool. *J. Mol. Biol.* **215,** 403–410 (1990).

797 22. Berglund, F. *et al.* Identification of 76 novel B1 metallo-β-lactamases through large-scale

798 screening of genomic and metagenomic data. *Microbiome* **5,** (2017).

799 23. Everard, A. *et al.* Cross-talk between Akkermansia muciniphila and intestinal epithelium controls

800 diet-induced obesity. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 9066–9071 (2013).

801 24. Sokol, H. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium

802 identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U. S. A.* **105,**

803 16731–16736 (2008).

804 25. Leski, T. A. *et al.* Multidrug-resistant tet(X)-containing hospital isolates in Sierra Leone. *Int. J.*

805   *Antimicrob. Agents* **42,** 83–86 (2013).

806 26. Oñate, F. P. *et al.* MSPminer: abundance-based reconstitution of microbial pan-genomes from

807   shotgun metagenomic data. *bioRxiv* 173203 (2018). doi:10.1101/173203

808 27. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat.*

809   *Biotechnol.* **32,** 834–841 (2014).

810 28. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).

811 29. de Smet, A. M. G. A. *et al.* Decontamination of the digestive tract and oropharynx in ICU patients.

812   *N. Engl. J. Med.* **360,** 20–31 (2009).

813 30. van Schaik, W. The human gut resistome. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370,** (2015).

814 31. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers.

815   *Nature* **500,** 541–546 (2013).

816 32. Pehrsson, E. C. *et al.* Interconnected microbiomes and resistomes in low-income human habitats.

817   *Nature* **533,** 212–216 (2016).

818 33. Léonard, F., Andremont, A., Leclerq, B., Labia, R. & Tancrède, C. Use of beta-lactamase-

819   producing anaerobes to prevent ceftriaxone from degrading intestinal resistance to colonization. *J.*

820   *Infect. Dis.* **160,** 274–280 (1989).

821 34. Bilinski, J. *et al.* Fecal Microbiota Transplantation in Patients with Blood Disorders Inhibits Gut

822   Colonization with Antibiotic-Resistant Bacteria: Results of a Prospective, Single-Center Study.

823   *Clin. Infect. Dis.* (2017). doi:10.1093/cid/cix252

824 35. Lupo, A., Coyne, S. & Berendonk, T. U. Origin and Evolution of Antibiotic Resistance: The

825   Common Mechanisms of Emergence and Spread in Water Bodies. *Front. Microbiol.* **3,** (2012).

826 36. Lagier, J.-C. *et al.* Culture of previously uncultured members of the human gut microbiota by

827   culturomics. *Nat. Microbiol.* **1,** 16203 (2016).

828 37. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than

829   sequence--a study of structural response in protein cores. *Proteins* **77,** 499–508 (2009).

830 38. Baquero, F., Tedim, A. P. & Coque, T. M. Antibiotic resistance shaping multi-level population

831   biology of bacteria. *Front. Microbiol.* **4,** 15 (2013).

832 39. Martínez, J. L., Coque, T. M. & Baquero, F. What is a resistance gene? Ranking risk in

833   resistomes. *Nat. Rev. Microbiol.* **13,** 116–123 (2015).

834 40. Van Boeckel, T. P. *et al.* Global antibiotic consumption 2000 to 2010: an analysis of national

835 pharmaceutical sales data. *Lancet Infect. Dis.* **14,** 742–750 (2014).

836 41. Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A. & Handelsman, J. Functional metagenomics

837 reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J.* **3,** 243–251 (2009).

838 42. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification.

839 *BMC Bioinformatics* **11,** 119 (2010).

840 43. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation

841 sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).

842 44. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different

843 taxonomic ranges. *Nucleic Acids Res.* **40,** D284-289 (2012).

844 45. Liu, Y.-Y. *et al.* Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals

845 and human beings in China: a microbiological and molecular biological study. *Lancet Infect. Dis.*

846 **16,** 161–168 (2016).

847 46. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity

848 searching. *Nucleic Acids Res.* **39,** W29-37 (2011).

849 47. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl.*

850 *Acad. Sci. U. S. A.* **85,** 2444–2448 (1988).

851 48. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

852 49. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments

853 using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).

854 50. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J.*

855 *Mol. Biol.* **292,** 195–202 (1999).

856 51. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol.*

857 *Biol.* **234,** 779–815 (1993).

858 52. Wallner, B. & Elofsson, A. Can correct protein models be identified? *Protein Sci. Publ. Protein*

859 *Soc.* **12,** 1073–1086 (2003).

860 53. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in

861 three-dimensional structures of proteins. *Nucleic Acids Res.* **35,** W407-410 (2007).

862   54. Ortiz, A. R., Strauss, C. E. M. & Olmea, O. MAMMOTH (matching molecular models obtained

863   from theory): an automated method for model comparison. *Protein Sci. Publ. Protein Soc.* **11,**

864   2606–2621 (2002).

865   55. Tibshirani, R. Regression shrinkage and selection via the lasso. *J R. Stat. Soc B* **58,** 267–288

866   (1996).

867   56. Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R. & Lin, C. J. A library for large linear

868   classification. *JMLR* **9,** 1871–1874

869   57. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre

870   for bacterial insertion sequences. *Nucleic Acids Res.* **34,** D32-36 (2006).

871   58. Guglielmini, J., Quintais, L., Garcillán-Barcia, M. P., de la Cruz, F. & Rocha, E. P. C. The

872   repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS*

873   *Genet.* **7,** e1002222 (2011).

874   59. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial

875   metagenomics. *PloS One* **7,** e30126 (2012).

876   60. Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the

877   human body. *Nature* **509,** 357–360 (2014).

878   61. Dray, S. & Legendre, P. Testing the species traits-environment relationships: the fourth-corner

879   problem revisited. *Ecology* **89,** 3400–3412 (2008).

880   62. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of

881   an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl. Environ.*

882   *Microbiol.* **63,** 2802–2813 (1997).

883   63. Suau, A. *et al.* Direct analysis of genes encoding 16S rRNA from complex communities reveals

884   many novel molecular species within the human gut. *Appl. Environ. Microbiol.* **65,** 4799–4807

885   (1999).

886   64. Pons, N. *et al.* METEOR - a plateform for quantitative metagenomic profiling of complex

887   ecosystems. in (2010).

888   65. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–

889   359 (2012).

890   66. Quereda, J. J. *et al.* Bacteriocin from epidemic Listeria strains alters the host intestinal microbiota

891   to favor infection. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 5706–5711 (2016).

892   67. Robert, P. & Escoufier, Y. A Unifying Tool for Linear Multivariate Statistical Methods: The RV-

893   Coefficient. *Appl. Stat.* **25,** 257 (1976).

894

Figure 1: Illustration of the concept of "Pairwise Comparative Modelling" (PCM) with a class A beta-lactamase (panel A). A1: class A beta-lactamase protein structure (4EWF) obtained from the PDB database. A2: A candidate protein (MC3.MG12.AS1.GP1.C14.G3 from *Faecalibacterium prausnitzii*) for class A beta-lactamase modelled with a reference class A beta-lactamase structural template. This protein had 26.5% amino acid identity with the closest reference class A beta-lactamase. A3: The same candidate protein (MC3.MG12.AS1.GP1.C14.G3) for class A beta-lactamase this time modelled with a negative reference template. The candidate MC3.MG12.AS1.GP1.C14.G3 was predicted to be a class A beta-lactamase with 100% confidence by our model and later found to be functional after gene synthesis. Panel B: Bar-plot of the activity of the synthesized pdARDs against antibiotics with respect to the degree of confidence of the prediction ("reference" meaning that the protein shares more ≥95% amino acid identity with a functionally proven ARD, "good" meaning a PCM score over 99% and a TmAlign Tm score ≥0.8, "fair" meaning a PCM score between 50% and 80%). Panel C: number of predictions of antibiotic resistance determinants from a 3.9 million gene catalogue of the intestinal microbiota[19] using PCM, BLASTP[21], ARG-ANNOT[7], Resfinder[9], DeepARG[10], Resfams[11] and the HMM-based method published by Berglund *et al.* for class B1 beta-lactamases[22]. Panel D: violin plot of the maximal identity observed with a reference ARD for candidates predicted as ARDs (blue violin, n=6,095) and those not predicted as ARDs (red violin, n=3,982). The point depicts the median. The width of the violins depicts the distribution of pdARDs according to their maximal identity with a reference ARD. See Supplementary Table 2 for details about candidates sharing at least 40% identity with reference ARDs but which were not predicted as ARDs.
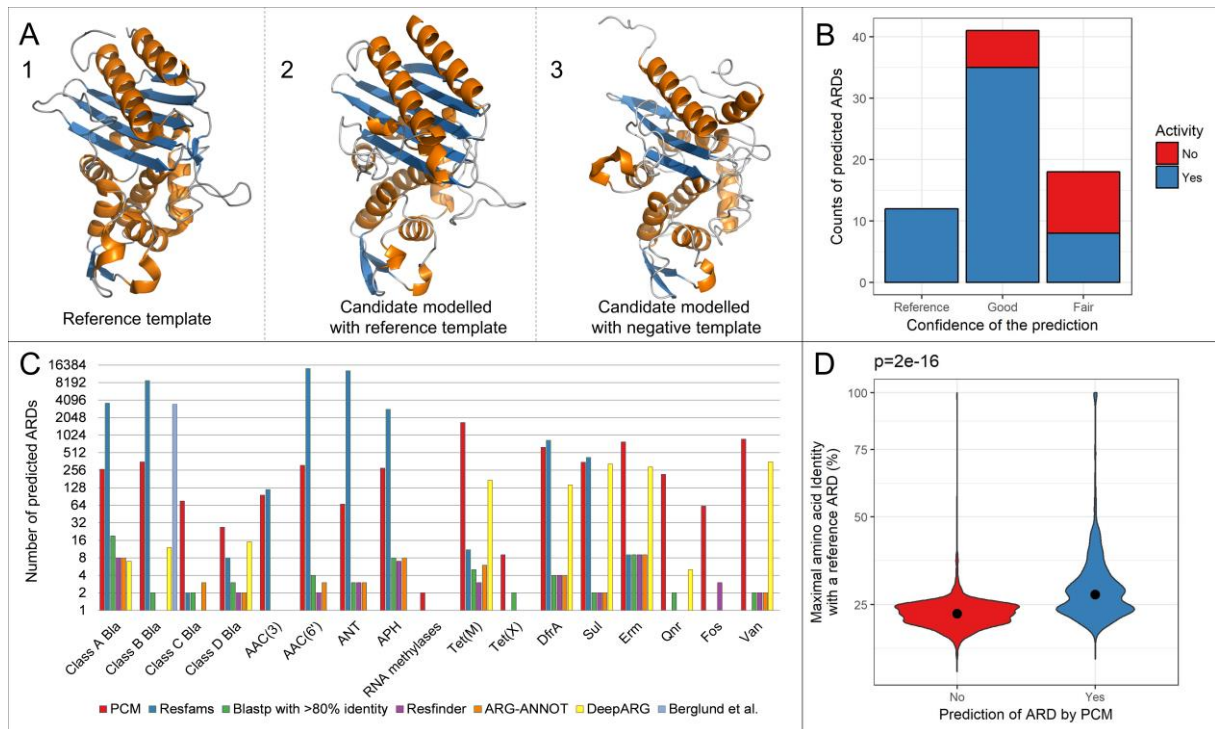
Bla: beta-lactamase; AAC: aminoglycoside acetylase; ANT: aminoglycoside nucleotidyl transferase; APH: aminoglycoside phosphotransferase; DfrA: type A dihydrofolate reductase; Sul: dihydropteroate synthase; Erm: erythromycin ribosome methylase; Qnr: quinolone resistance; Fos: fosfomycin resistance (Fos); Van: D-Ala – D-Lac/Ser ligase (vancomycin resistance).

922 **Figure 2:** Mobile genetic elements (MGE) and predicted antibiotic resistance determinants (pdARDs).

923 (A) Distribution of the sizes of the metagenomics unit (MGU) where an antibiotic resistance determinant

924 was predicted with respect to the colocation of MGE-associated genes. The vertical line depicts the

925 assumed gene size threshold above which MGUs are considered as partial chromosomes referred as

926 metagenomic species (MGS)[19]. (B) Bar plot of the categories of metagenomic species pangenomes

927 (MSPs)[26] assigned to MGE – associated genes[27] and pdARDs. (C) Proportion of pdARDs co-locating

928 with MGE-associated genes with respect to their phylum. (D) Proportion of pdARDs co-locating with

929 MGE-associated genes according to the pdARD family. Of note, the AAC(2') and 16S RNA methylases

930 only included 3 and 2 pdARDs, respectively and were accordingly not depicted in this panel.

931
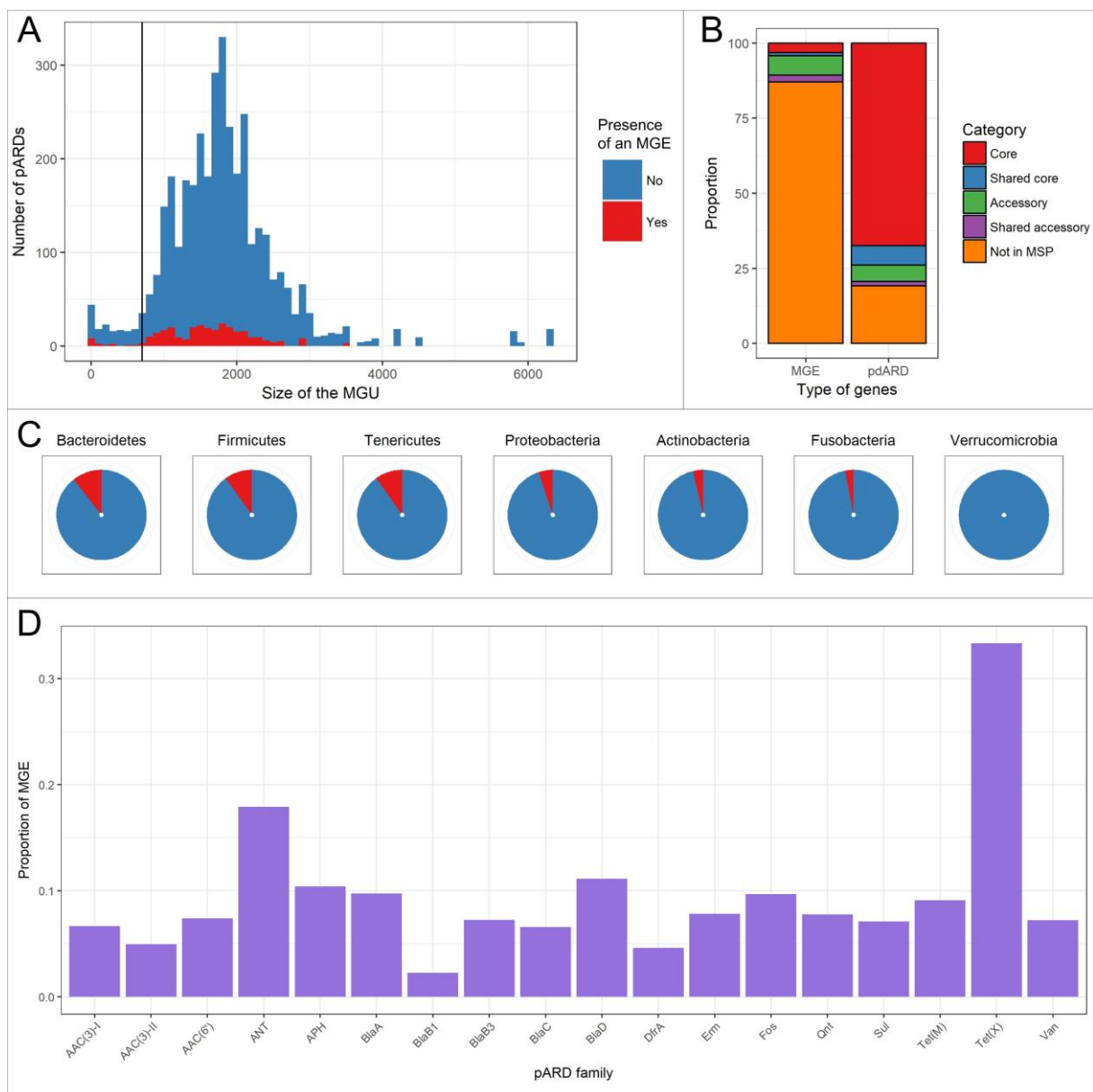
Figure 3: Association between resistotypes, enterotypes, metagenomics species (MGS) and pdARDs profiles in the 663 individuals from the MetaHIT cohort. A) inertia shared between pdARDs profiles and microbiota composition as function of bioinformatics methods. We assessed how gut microbiota beta diversity inertia was connected to the abundance of pdARDs. Co-inertia using RV coefficient was analysed to detect significant co-structure between datasets[67], meaning that different sets of variables (e.g. microbial genera abundance and ARDs profiles) were not independent and shared a fraction of inertia. Monte- Carlo tests were used to confirm observed relations between different datasets, assuming a p-value < 0.05. B) Samples proportions for each resistotype depicted as function of enterotypes using the PCM method. C) and D) Association between pdARDs gene profile and gut microbiota composition using co-inertia analysis with respect to their enterotypes and pdARDs families (C), and to their resistotypes and MGS relative abundance (D). A taxonomical correspondence for each MGS number can be found in the original paper[19]. Briefly, all MGS were Firmicutes with the exception of MGS:164 and MGS:445 (both Bacteroidetes).
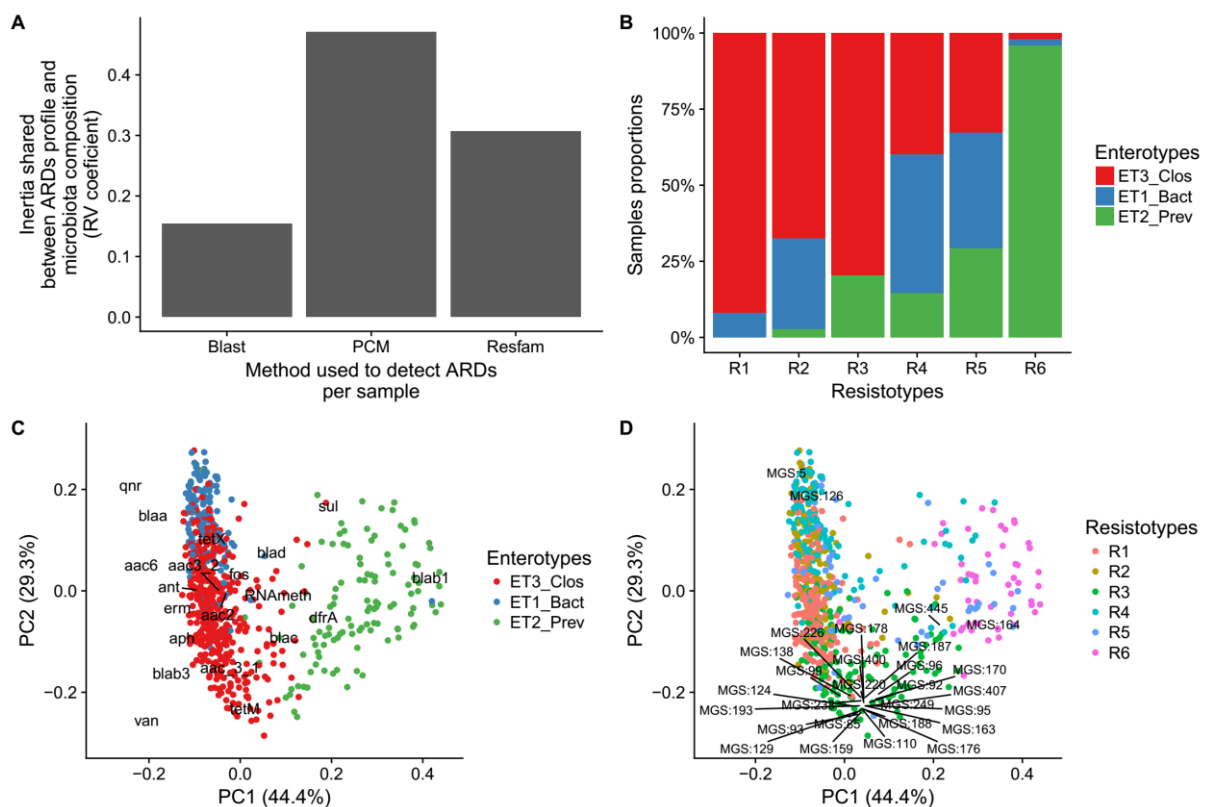
Figure 4: (A) Gene richness and relative abundance of predicted antibiotic resistance determinants (pdARDs) in the MetaHIT cohort (n=663). (B) Gene richness and relative abundance of pdARDs in our cohort of subjects with no recent antibiotic exposure (n=44). (C) Gene richness and relative abundance of pdARDs in our cohort of subjects with regards to their antibiotic exposure (n=102 samples). (D) and (E) Boxplots superimposed by dot plots of the comparisons of the relative abundance of all pdARDs and gene richness, respectively, between the various groups differing by their exposure to antibiotics. Hospitalization: n=15, Wilcoxon paired test. Short high dose exposure: n=10, Wilcoxon paired test. Chronic exposure: n=31 for patients not exposed to antibiotics, n=30 for patients chronically exposed to antibiotics, Wilcoxon unpaired test.
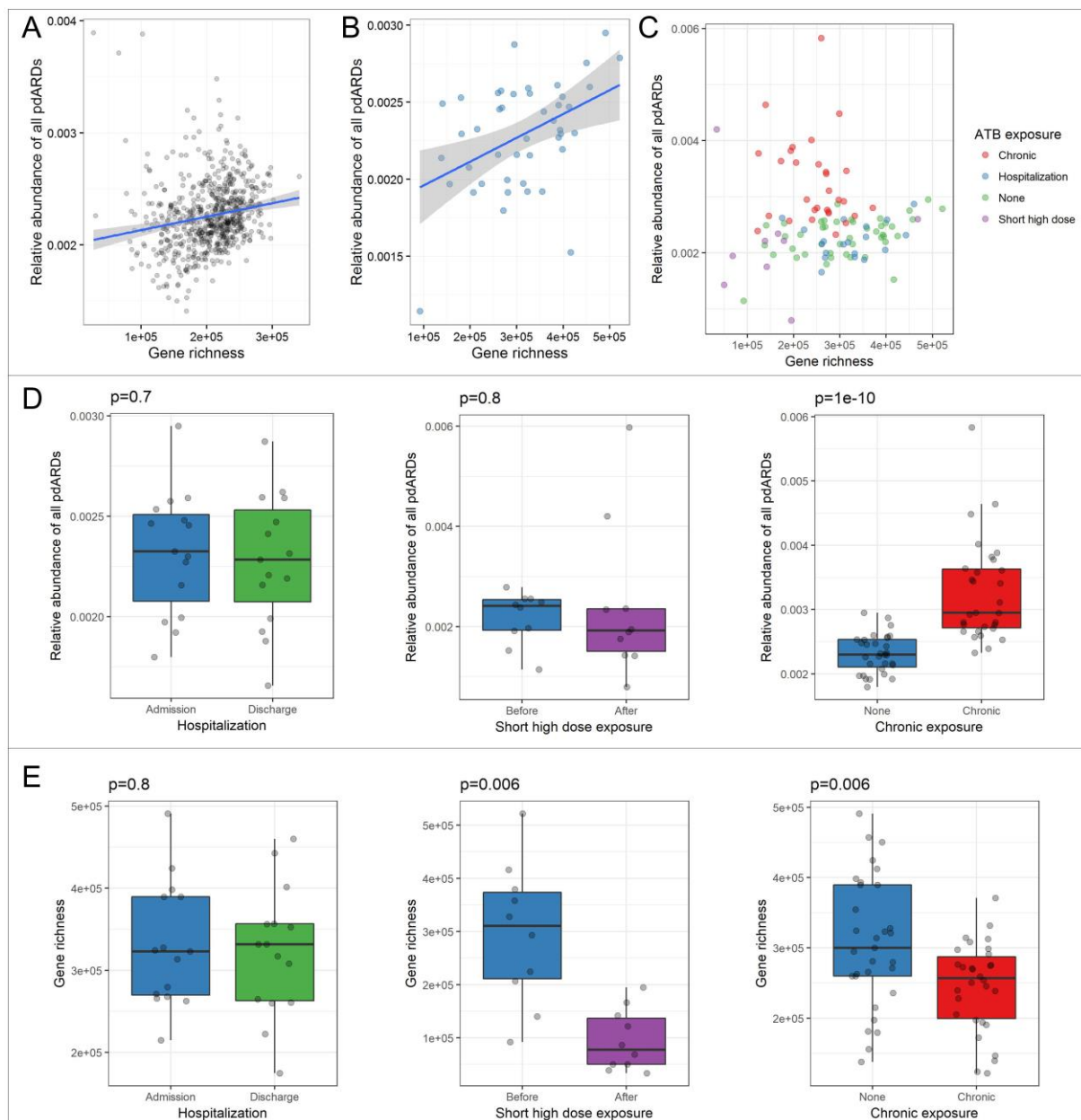
958   ATB: antibiotics. The shaded grey area depicts the 95% confidence interval around the blue, linear

959   regression line. For boxplots, the lower, central and upper hinges correspond to the first, second

960   (median) and third quartiles. The upper and lower whiskers respectively correspond to the higher and

961   lower values at 1.5*IQR from the hinge (where IQR is the inter-quartile range, or distance between the

962   first and third quartiles).

963   **Table 1**: Summary of the predictions of antibiotic resistance determinants (ARDs) from a 3.9 million

964   gene catalogue of the intestinal microbiota[19] and of gene synthesis results.

Antibiotic resistance class	Number of references	Number of candidates	Number of predictions	Rate ARD predictions/candidates (%)	Tested (%)	N functional (%)	N not functional (%)
16S rRNA methylase	17	4	2	50,0	0 (0%)	NA	NA
AAC(2')	5	15	3	20,0	0 (0%)	NA	NA
AAC(3)-I	7	53	15	28,3	2 (13.3%)	2 (100%)	0 (0%)
AAC(3)-II	12	81	81	100	5 (6.2%)	5 (100%)	0 (0%)
AAC(6')	36	1191	312	26,2	8 (2.6%)	6 (75%)	2 (25%)
ANT	29	158	67	42,4	3 (4.5%)	3 (100%)	0 (0%)
APH	30	430	279	64,9	4 (1.4%)	3 (75%)	1 (25%)
Class A beta-lactamase	682	402	267	66,4	14 (5.2%)	9 (64.3%)	5 (35.7%)
Class B1-B2 beta-lactamase	150	554	134	24,2	8 (6.0%)	6 (75%)	2 (25%)
Class B3 beta-lactamase	31	493	221	44,8	7 (3.2%)	5 (71.4%)	2 (28.6%)
Class C beta-lactamase	56	373	76	20,4	4 (5.3%)	4 (100%)	0 (0%)
Class D beta-lactamase	248	76	27	35,5	2 (7.4%)	2 (100%)	0 (0%)
DfrA	35	632	632	100	0 (0%)	NA	NA
Erm	58	873	781	89,5	0 (0%)	NA	NA
Fos	34	84	62	73,8	0 (0%)	NA	NA
Qnr	66	272	219	80,5	0 (0%)	NA	NA
Sul	33	357	353	98,9	0 (0%)	NA	NA
Tet(M)	72	2824	1682	59,6	13 (0.8%)	9 (69.2%)	4 (30.8%)
Tet(X)	12	42	9	21,4	1 (11.1%)	1 (100%)	0 (0%)
Van ligase	16	1163	873	75,1	0 (0%)	NA	NA

965