

1 **Flanker: a tool for comparative genomics of gene flanking regions**

2 William Matlock<sup>1\*</sup>, Samuel Lipworth<sup>1,2\*</sup>, Bede Constantinides<sup>1,3</sup>, Timothy E.A. Peto<sup>1,2,3,4</sup>, A.  
3 Sarah Walker<sup>1,3,4</sup>, Derrick Crook<sup>1,2,3,4</sup>, Susan Hopkins<sup>5</sup>, Liam P. Shaw<sup>6^</sup> and Nicole  
4 Stoesser<sup>1,2,3^</sup>

5 \* Joint first authors

6 ^ Joint senior authors

7 <sup>1</sup> Nuffield Department of Medicine, University of Oxford, UK

8 <sup>2</sup> Oxford University Hospitals NHS Foundation Trust, Oxford, UK

9 <sup>3</sup> NIHR Health Protection Research Unit in Healthcare Associated Infections and  
10 Antimicrobial Resistance at University of Oxford in partnership with Public Health England,  
11 Oxford, United Kingdom

12 <sup>4</sup> NIHR Oxford Biomedical Research Centre, Oxford, UK

13 <sup>5</sup> National Infection Service, Public Health England, Colindale, London, UK

14 <sup>6</sup> Department of Zoology, University of Oxford, UK

15 **Corresponding author:** william.matlock@ndm.ox.ac.uk

16 **Keywords:** mobile genetic elements (MGEs); plasmids; whole genome sequencing;  
17 bioinformatics, antimicrobial resistance (AMR)

18

19    **Abstract**

20    **Analysing the flanking sequences surrounding genes of interest is often highly relevant to**  
21    **understanding the role of mobile genetic elements (MGEs) in horizontal gene transfer,**  
22    **particular for antimicrobial resistance genes. Here, we present Flanker, a Python package**  
23    **which performs alignment-free clustering of gene flanking sequences in a consistent**  
24    **format, allowing investigation of MGEs without prior knowledge of their structure. These**  
25    **clusters, known as 'flank patterns', are based on Mash distances, allowing for easy**  
26    **comparison of similarity across sequences. Additionally, Flanker can be flexibly**  
27    **parameterised to finetune outputs by characterising upstream and downstream regions**  
28    **separately and investigating variable lengths of flanking sequence. We apply Flanker to**  
29    **two recent datasets describing plasmid-associated carriage of important carbapenemase**  
30    **genes (*blaOXA-48* and *blaKPC-2/3*) and show that it successfully identifies distinct clusters**  
31    **of flank patterns, including both known and previously uncharacterised structural**  
32    **variants. For example, Flanker identified four *Tn4401* profiles that could not be sufficiently**  
33    **characterised using TETyper or MobileElementFinder, demonstrating the utility of Flanker**  
34    **for flanking gene characterisation. Similarly, using a large (n=226) European isolate**  
35    **dataset, we confirm findings from a previous smaller study demonstrating association**  
36    **between *Tn1999.2* and *blaOXA-48* upregulation and demonstrate 17 flank patterns**  
37    **(compared to the 5 previously identified). More generally the demonstration in this study**  
38    **that flank patterns are associated with geographical regions and antibiotic**  
39    **susceptibility phenotypes suggests that they may be useful as epidemiological markers.**  
40    **Flanker is freely available under an MIT license at <https://github.com/wtmatlock/flanker>.**

41

42    **Data Summary**

43 NCBI accession numbers for all sequencing data used in this study is provided in Supplementary  
44 Table 1. The analysis performed in this manuscript can be reproduced in a binder environment  
45 provided on the Flanker Github page (<https://github.com/wtmatlock/flanker>).

46 **Introduction**

47 The increasing incidence antimicrobial resistance (AMR) in clinical isolates poses a threat to all areas  
48 of medicine [1–3]. AMR genes (ARGs) are found in a diverse range of genetic contexts, bacterial  
49 species, and in both clinical and non-clinical environments (e.g., agricultural, refuse and natural  
50 ecosystems) [4–7]. However, the mechanisms underpinning the dissemination of many ARGs  
51 between these reservoirs remain poorly understood, limiting the efficacy of surveillance and the  
52 ability to design effective interventions. Usually, ARGs are spread vertically, either via chromosomal  
53 integration or stable association of a plasmid within a clonal lineage, or by horizontal gene transfer  
54 (HGT) through mobile genetic elements (MGEs) e.g., transposons or plasmids [8]. HGT can  
55 accelerate the rate of ARG acquisition, both within and across species [9–11].

56 The epidemiology of ARGs can therefore involve multiple levels, from clonal spread to MGEs. There  
57 are many existing software tools to facilitate epidemiological study of bacterial strains [12–16], whole  
58 plasmids [17, 18], and smaller MGEs [19, 20]. Several tools and databases exist for the annotation of  
59 non-plasmid MGEs such as insertion sequences (ISs) and transposons [19, 20], but all rely on  
60 comparisons to reference sequences, so are limited to known diversity. Reference-free tools for  
61 analysing MGE diversity would therefore be a useful addition. Here we describe a Flanker, a simple,  
62 reference-free tool to investigate MGEs by analysing the flanking sequences of ARGs.

63 The flanking sequences (hereafter, ‘flanks’) around an ARG that has been mobilised horizontally may  
64 act as signatures of relevant MGEs and support epidemiological analyses. However, these flanks can  
65 contain a great deal of structural variation due to their evolutionary history. Where a single known  
66 MGE is under investigation, it is possible to specifically type this element (for example, using  
67 TETyper [19]) or align flanks against a known ancestral form after the removal of later structural  
68 variation [21]. However, often multiple structures may be involved. This is particularly true for ARGs

69 which move frequently on a variety of MGEs. Studies of different ARGs often choose different *ad*  
70 *hoc* approaches to extract flanks and cluster genetic structures. Examples include hierarchical  
71 clustering of isolates carrying an ARG based on short-read coverage of known ARG-carrying contigs  
72 [22], assigning assembled contigs into 'clustering groups' based on gene presence and synteny [23] or  
73 iterative 'splitting' of flanks based on pairwise nucleotide BLAST identity [24]. A consistent and  
74 simple approach for this task would not only avoid repeated method development, but also aid  
75 comparison between methods developed for specific ARGs.

76 To address this problem, we developed Flanker, a pipeline to analyse the regions around a given ARG  
77 in a consistent manner. Flanker flexibly extracts the flanks of a specified gene from a dataset of  
78 contigs, then clusters these sequences using Mash distances to identify consistent structures [25].  
79 Flanker is available as a documented Python and Bioconda package released under the MIT open-  
80 source license. Source code is deposited at <https://github.com/wtmatlock/flanker> and documentation  
81 at <https://flanker.readthedocs.io/en/latest/>.

82

### 83 **Methods**

#### 84 Flanker

85 The Flanker package contains two basic modules: the first extracts a region of length  $w$  around an  
86 annotated gene of interest, and the second clusters such regions based on a user-defined Mash  
87 distance threshold (default --threshold 0.001, Fig. 1a). Within each FASTA/multi-FASTA format  
88 input file, the location of the gene of interest is first determined using the Abricate annotation tool  
89 [31]. Flanks around the gene (optionally including the gene itself to enable complete alignments with  
90 --include\_gene) are then extracted and written to a FASTA format file using BioPython [39]. Flanker  
91 gives users the option to either extract flanks using a single window (defined by length in base-pairs  
92 [bp]) or multiple windows from a start position (--window) to an end position (--wstop) in fixed  
93 increments (--wstep). Flanks may be extracted from upstream, downstream or on both sides of the  
94 gene of interest (--flank). Corrections are also made for circularised genomes where the gene occurs

95 close to the beginning or end of the sequence (--circ mode) and for genes found on both positive and  
96 negative strands. The clustering module groups flanks of user-defined sequence lengths together  
97 based on a user-defined Mash [25] distance threshold (--threshold) of user-defined sequence lengths.

98 In default mode (--mode default), Flanker considers multiple gene queries in turn. In multi-allelic  
99 mode (--mode mm), Flanker considers all genes in the list for each window (Fig. 1b). Multiple genes  
100 can be queried by either a space-delimited list in the command line (--gene geneA geneB), or a  
101 newline-delimited file with the list of genes option (--list\_of\_genes). A supplementary module ‘salami  
102 mode’ (--mode sm) is provided to allow comparison of non-contiguous blocks from a start point (--  
103 window), step size (--wstep) and end point (--wstop) (Fig. 1c).

104 Datasets

105 To validate Flanker, demonstrate its application and provide a comparison with existing tools, we  
106 used two recent datasets of complete plasmids (derived from hybrid long-/short-read assemblies)  
107 containing carbapenemase genes of clinical importance [23, 26]. The first dataset comprised 51  
108 complete *bla*<sub>OXA-48</sub>-harbouring plasmids; 42/51 came from carbapenem-resistant *Escherichia coli* and  
109 *Klebsiella pneumoniae* isolates from patients in the Netherlands [26] and 9/51 from EuSCAPE (a  
110 European surveillance programme investigating carbapenem resistance in Enterobacterales) [23]. The  
111 second dataset comprised 50 *bla*<sub>KPC-2</sub> or *bla*<sub>KPC-3</sub>-(*Klebsiella pneumoniae* carbapenemase)-harbouring  
112 plasmids in carbapenem-resistant *K. pneumoniae* isolated from the Netherlands [26] (8/50) and as part  
113 of the EuSCAPE study (42/50)[23]. The EuSCAPE dataset [23, 27] additionally contains a large  
114 collection of short-read sequencing data for *Klebsiella* spp. isolates alongside meropenem  
115 susceptibility data. This was used to demonstrate additional possible epidemiological applications of  
116 the Flanker tool by evaluating whether specific flank patterns were more likely to be associated with  
117 phenotypic meropenem resistance.

118 Mash distances

119 Pair-wise distances between flanks were calculated using Mash (version 2.2.2) [25]. Mash reduces  
120 sequences to a fixed-length MinHash sketch, which is used to estimate the Jaccard distance between

121 k-mer content. It also gives the Mash distance, which ranges from 0 (~identical sequences) to 1  
122 (~completely dissimilar sequences). We used the default Mash parameters in all analyses. The Mash  
123 distance was developed to approximate the rate of sequence mutation between genomes under a  
124 simple evolutionary model, and explicitly does not model more complex processes. We use it here for  
125 fast alignment-free clustering of sequences and do not draw any direct conclusions about evolution  
126 from pairwise comparisons.

127 Clustering

128 To cluster the flanks, Flanker generates an adjacency matrix weighted by Mash distances. It then  
129 thresholds this matrix to retain edges weighted less than or equal to the defined threshold. This is then  
130 used to construct a graph using the Python NetworkX library [28] and clusters are defined using the  
131 nx.connected\_components function, which is analogous to single linkage. This is a similar  
132 methodology to that used by the Assembly De-replicator tool [29] (from which Flanker re-uses  
133 several functions). However, Flanker aims to assign all flanks to a cluster rather than to deduplicate  
134 by cluster.

135 Cluster validation

136 We validated the output of flanking sequence-based clustering using a PERMANOVA test,  
137 implemented with the adonis function from the Vegan package (version 2.7.5) [30] in R. Only flanks  
138 in clusters of at least two members were considered; 42/51 and 48/50 of *bla*<sub>OXA-48</sub> and *bla*<sub>KPC-2/3</sub> flanks,  
139 respectively. The formula used was Mash dist ~ cluster, with the ‘euclidean’ method and 999  
140 permutations.

141 Comparison to existing methods/application

142 We compared the classifications of TETyper (v1.1) [19] and MEFinder (v1.0.3) [20] to those  
143 produced by Flanker for 500bp and 5000bp flanks around *bla*<sub>KPC-2/3</sub> genes. TETyper was run using the  
144 –threads 8 and --assemblies options with the Tn4401 reference and SNP/structural profiles provided  
145 in the package and MEFinder was run in Abricate[31] using the –mincov 10 option. For comparisons  
146 of the proportions of resistant isolates per flank pattern (denoted FP), isolates were classified as

147 resistant or sensitive using the European Committee on Antimicrobial Susceptibility Testing  
148 (EUCAST) breakpoint for meropenem (>8mg/L) [32].

149 **Data visualisation**

150 All figures were made using Biorender (<https://biorender.com>) and the R packages ggplot2 (v3.3.0)  
151 [33], ggenes (v0.4.0) [34] and ggtree (v2.4.1) [35] Prokka (v1.14.6) [36] was used to annotate  
152 Flanker output. MashTree (v1.2.0) [37] was used to construct a visual representation of Mash distances  
153 between whole plasmid genomes. Plasmidfinder was used to detect the presence/absence of plasmid  
154 types using Abricate (version 1.01) with --mincov 80 and --minid 80 [38]. Galileo AMR  
155 (<https://galileoamr.arcbio.com/mara/>) was used to visualise the transposon variants. Figures can be  
156 reproduced using the code in the GitHub repository (<https://github.com/wtmatlock/flanker>).

157

158 **Results**

159 **Clustering validation and comparison with TETyper/MEFinder**

160 The clustering mode was validated numerically with a PERMANOVA test (Mash dist ~ cluster:  
161 *bla*<sub>OXA-48</sub> *p*-value < 0.001, *bla*<sub>KPC2/3</sub> *p*-value < 0.001; see Methods). Figures 2 and 3 also provide a  
162 visual comparison of an alignment of genes ('Gene graphical representation' panel) to the Flank  
163 pattern.

164 Of the two existing tools we compared in evaluating the flanks around *bla*<sub>KPC2/3</sub>, TETyper was by far  
165 the slowest (1172 seconds [s]), whereas MEFinder, run in Abricate, and Flanker took 7s and 11s,  
166 respectively (benchmarked on 5000bp upstream flanks on a cluster with Intel Skylake 2.6GHz chips).  
167 MEFinder was able to detect *Tn4401* but could not provide any further structural resolution and was  
168 unable to classify 6/50 (12%) 500bp and 1/50 (2%) 5000bp flanks. TETyper structural profiles were  
169 consistent with Flanker when analysing 500 and 5000bp upstream regions (Figure 3), though Flanker  
170 split a group of six isolates with the TETyper structural profile 1-7127|7202-10006 into four groups  
171 (Table S2). To map our FPs to the established nomenclature, we additionally compared the output of

172 Flanker to that of TETyper when the latter was given the entire *Tn4401* sequence (i.e., by evaluating  
173 the typical 7,200bp *Tn4401*-associated flank upstream of *bla<sub>KPC</sub>*). Flanker and TETyper classifications  
174 of *Tn4401* regions were broadly consistent (Table S2), though this analysis demonstrated the potential  
175 benefit of the reference-free approach of Flanker which showed that four non-*Tn4401* structural  
176 profiles ('unknown' in TETyper) were distinct from each other. In addition, TETyper classified 3  
177 flanks as *Tn4401\_truncC-1*, whereas Flanker resolved this cluster into two distinct groups (Table S2).

178

179 Application to plasmids carrying *bla<sub>OXA-48</sub>*

180 The carbapenemase gene *bla<sub>OXA-48</sub>* has been shown to be disseminated by *Tn1999*-associated  
181 structures (~5kb, see detailed review in [40]) nested in L/M-type plasmids, and as part of an IS1R-  
182 associated composite transposon containing *bla<sub>OXA-48</sub>* and part of *Tn1999*, namely *Tn6237* (~21.9kb),  
183 that has been implicated in the chromosomal integration of *bla<sub>OXA-48</sub>* [27, 41]. It has been recently  
184 demonstrated that most *bla<sub>OXA-48</sub>*-like genes in clinical isolates in Europe are carried on highly similar  
185 L/M(pOXA-48)-type plasmids, with evidence of both horizontal and vertical transmission across a  
186 diverse set of sequence types [23]. Whilst *Tn1999*-like flanking regions are relatively well  
187 characterised [40], in this example we chose an initial arbitrary upstream window of 5000bp to  
188 simulate a scenario in which there is no prior knowledge. Inspection of a plot of window clusters (i.e.,  
189 as shown in the 'Flankergram' in Fig. 2) demonstrates that Flanker output allows the empirical  
190 identification of the position ~2200bp upstream of *bla<sub>OXA-48</sub>* as an important point of structural  
191 divergence without requiring annotation (as shown at ~2200 along the x-axis, where the window  
192 cluster colour schemes diverge), corresponding to the edge of *Tn1999* at its expected position.

193 Using complete plasmids from the Netherlands [26]/EUSCAPE [23] hybrid assembly datasets,  
194 Flanker identified 17 distinct FPs in the 2200bp upstream sequence of *bla<sub>OXA-48</sub>* of which seven  
195 occurred in L/M(pOXA-48)-type plasmids (Fig. 2, Table S3). To investigate the association of  
196 phenotypic carbapenem resistance with *bla<sub>OXA-48</sub>* FPs, we created a Mash sketch using one randomly  
197 chosen representative per group and screened an Illumina sequenced collection of European

198 carbapenemase-resistant *Klebsiella* isolates [27] for containment (n=425) (Mash screen, assigning FP  
199 based on the top hit [median identity = 1.00; range: 0.97-1.00]). Two FPs (FP6 and FP16) accounted  
200 for 338/425 (80%) of isolates; both were widely distributed across Europe. Of the 226 isolates with  
201 meropenem susceptibility data available, those belonging to FP6 were proportionally more  
202 meropenem-resistant compared to FP16 (70/135 [52%] vs. 6/44 [14%], exact *p*-value<0.001; Fig.3).  
203 Annotation (using Galileo AMR; see methods) of these revealed that whereas FP16 contains Tn1999,  
204 FP6 contains Tn1999.2, which has been previously described as creating a strong promoter which  
205 produces 2-fold higher enzymatic activity [42].

206 Application to plasmids carrying bla<sub>KPC-2/3</sub>

207 David et al. showed that bla<sub>KPC-2/3</sub> genes have been disseminated in European *K. pneumoniae* clinical  
208 isolates via a diverse collection of plasmids in association with a dominant clonal lineage, ST258/512,  
209 which accounted for 230/312 (74%) of bla<sub>KPC</sub>-associated isolates in the EuSCAPE collection [23].  
210 bla<sub>KPC</sub> has largely been associated with variants of a ~10kb transposon, Tn4401 [43, 44]. From the  
211 combined EuSCAPE [23] and Dutch CPE collection [26] of 50 hybrid assembled KPC-containing  
212 plasmids, Flanker identified 8 distinct FPs over a 7200bp window upstream of bla<sub>KPC-2/3</sub> (Fig. 4; Table  
213 S2). This window length was chosen to capture the entire Tn4401 sequence upstream of bla<sub>KPC</sub>.

214  
215 Considering Mash containment of the 8 representative FPs within the EuSCAPE short read assemblies  
216 dataset, 346/442 (78%) belonged to FP1 (corresponding to isoform Tn4401a). Whilst FP1 was widely  
217 distributed across Europe, FP2 (corresponding to Tn4401\_truncC) and FP7 (corresponding to  
218 Tn4401d) appeared more geographically restricted: FP2 to Spain (5/5, 100%) and FP7 to Israel  
219 (19/59, 32%) and Portugal (34/59, 58%) with isolates also found in Poland and Germany (n=2 each)  
220 and Italy and Austria (n=1, Table S3). Of the 442 short read assemblies, 274 had meropenem MIC  
221 data available for analysis. There was no evidence of a difference in the proportion of isolates  
222 resistant to meropenem between FP1 and FP7 (202/238 [85%] vs 23/25 [92%], exact *p*-value=0.5,

223 Table S4), though there was incomplete susceptibility data for isolates from both groups (108/346  
224 [31%] for FP1 and 38/63 [60%] for FP7).

225 **Discussion**

226 We present Flanker, a fast and flexible Python package for analysing gene flanking sequences. We  
227 anticipate that this kind of analysis will become more common as the number of complete reference-  
228 grade, bacterial assemblies increase. Our analysis of data from the EuSCAPE project suggests that  
229 flank patterns (FPs) might be useful epidemiological markers when evaluating geographical  
230 associations of sequences. Additionally, we validated findings of a small (n=7) PCR-based study on a  
231 large (n=226) European dataset, confirming an association between Tn1999.2 and increased  
232 meropenem resistance. A key advantage compared to existing tools is that there is no reliance on  
233 reference sequences or prior knowledge. Despite analysing only a relatively small number (n=50) of  
234 complete *bla<sub>KPC</sub>* containing plasmids, there were four distinct FPs which TETyper classified as  
235 'unknown' because their profiles had not been previously characterised. Similarly, we identified 17  
236 FPs associated with *bla<sub>OXA-48</sub>* in contrast to the five structural variants of Tn1999 currently described  
237 in the literature.

238

239 TETyper works well when alleles/structural variants are known but can only classify a single  
240 transposon type at a time and requires manual curation when this is not the case. The observed  
241 diversity of flanking sequences is likely to continue to increase and manual curation of naming  
242 schemes will be arduous to maintain. MEFinder on the other hand is a quick screening tool which can  
243 search a large library of known mobile elements but lacks sequence level resolution. Whilst Flanker  
244 overcomes these challenges, users may need to perform downstream analysis to interpret its output.  
245 We hope that Flanker will be complementary to these and other similar existing tools by reducing the  
246 dimensionality of large datasets and identifying smaller groups of sequences to focus on in detail.  
247 Though we have developed Flanker for ARGs, Abricate allows use of custom databases meaning any

248 desired genes of interest could be analysed. Accurate outputs from Flanker will be dependent on the  
249 quality of input assemblies, and on the correct annotation of the gene of interest.

250 In summary, we present Flanker, a tool for comparative genomics of gene flanking regions which  
251 integrates several existing tools (Abriicate, Biopython, NetworkX) in a convenient package with a  
252 simple command-line interface.

253

254 **Authors and contributors**

255 Contributions have been attributed by the CRediT system as follows:

256 Conceptualisation: WM, SL, LPS, NS

257 Methodology: WM, SL

258 Software: WM, SL, BC

259 Validation: WM, SL

260 Formal Analysis: WM, SL

261 Investigation: WM, SL

262 Resources: DWC, TP, ASW, NS

263 Data Curation: SL, WM

264 Writing - Original Draft Preparation: SL, WM, LPS, NS

265 Writing - Review and Editing: SL, WM, LS, BC, NS, DC, TP, ASW

266 Visualisation: SL, WM

267 Supervision: LPS, NS, TP, ASW, DC

268 Project Administration: SL, WM, NS, LPS

269 Funding: TP, DC, ASW, NS

270 **Data availability**

271 Accessions for the plasmid sequences, and MEFinder and TETyper outputs are provided in Table S1.

272 **Conflicts of Interest**

273 The authors have no conflicts of interest.

274 **Funding Information**

275 WM is supported by a scholarship from the Medical Research Foundation National PhD Training  
276 Programme in Antimicrobial Resistance Research (MRF-145-0004-TPG-AVISO). SL is an MRC  
277 Clinical Research Training Fellow (MR/T001151/1). LPS is a Sir Henry Wellcome Postdoctoral  
278 Fellow (220422/Z/20/Z). ASW and TP are NIHR Senior Investigators. The computational aspects of  
279 this research were funded from the NIHR Oxford BRC with additional support from the Wellcome  
280 Trust Core Award Grant Number 203141/Z/16/Z. The views expressed are those of the author(s) and  
281 not necessarily those of the NHS, the NIHR or the Department of Health. The research was supported  
282 by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare  
283 Associated Infections and Antimicrobial Resistance (NIHR200915) at the University of Oxford in  
284 partnership with Public Health England (PHE) and by Oxford NIHR Biomedical Research Centre.

285 **Acknowledgements**

286 The authors thank the EuSCAPE and Dutch CPE surveillance groups for making their data publicly  
287 available.

288 **Figure Descriptions**

289 **Figure 1: Schematic of Flanker's modes and parameters.** (a) Flanker uses Abricate to annotate the  
290 gene of interest in input sequences and outputs associated flanking sequences, optionally clustering (-  
291 cl) these on a user defined Mash distance threshold. It can take linear or circularised sequences. (b) In  
292 this example, genes *geneA* and *geneB* have been queried (-g *geneA* *geneB*), and only the upstream  
293 flank is desired (-f upstream). The top single black arrow represents choosing a single window of  
294 length 3000bp (-w 3000), whereas the bottom three black arrows represent stepping in 1000bp

295 windows from 0bp to 3000bp (-w 0 -wstep 1000 -wstop 3000). The default mode (-m default) extracts  
296 flanks for all annotated alleles separately, but the multi-allelic mode (-m mm) extracts flanks for all  
297 alleles in parallel. (c) Flanker has a supplementary salami mode (-m sm), which outputs non-  
298 contiguous blocks of sequence with a start point, step size, and end point (-w 0 -wstep 1000 -wstop  
299 3000), represented by the three black arrows.

300 **Figure 2: Flanking regions 5000bp upstream of *bla*<sub>OXA-48</sub> carrying plasmids from *Klebsiella*  
301 *pneumoniae* isolates.** The ‘Tree’ panel is a Neighbour-Joining tree constructed from Mash distances  
302 between complete sequences of plasmids carrying the *bla*<sub>OXA-48</sub> gene. The second panel indicates the  
303 presence/absence of a L/M(pOXA-48)-type plasmid. The ‘Gene Graphical Representation’ panel  
304 schematically represents coding regions in the 5000bp sequence upstream of the *bla*<sub>OXA-48</sub> gene, which  
305 is shown in red. Other genes are coloured according to the flank pattern which considers the overall  
306 pattern of all 100bp window clusters up to 2200bp (the approximate upstream limit of Tn1999). The  
307 “Flankergram” shows window clusters of all groups over each 100bp window between 0 and 5000bp.  
308 The dotted line at 2200bp indicates the approximate point of upstream divergence between several  
309 flank patterns. The ‘MLST’ panel shows *K. pneumoniae* multi-locus sequence types (MLSTs), with  
310 those occurring once labelled ‘other’. FPs are numbered in ascending order according to abundance in  
311 the hybrid assemblies. Data used to make this figure came from the Dutch CPE surveillance and  
312 EUSCAPE hybrid assembly datasets.

313

314 **Figure 3: Flanking regions 7200bp upstream of *bla*<sub>KPC-2/3</sub> carrying plasmids from *Klebsiella*  
315 *pneumoniae* isolates.** The ‘Tree’ panel is a Neighbour-Joining tree constructed from Mash distances  
316 between complete sequences of plasmids carrying the *bla*<sub>KPC-2/3</sub> gene. The next three panels indicate  
317 the presence/absence of FIB(pQ1I)-, FII(pKP91)-, and FIB(Kpn3)-type plasmids. The ‘Gene’ column  
318 indicates which *bla*<sub>KPC</sub> allele (2 or 3) is present. The ‘Gene Graphical Representation’ panel  
319 schematically represents coding regions in the 7200bp sequence region upstream of the *bla*<sub>KPC-2/3</sub>  
320 gene, which is shown in red. Other genes are coloured according to the flank pattern, which here takes  
321 into account the overall pattern of all 100bp window groups (shown in the “Flankergram” panel) over

322 the full 7200bp region upstream of *bla<sub>KPC-2/3</sub>*. The “Flankergram” shows window clusters over each  
323 100bp window between 0 and 7200bp. The ‘MLST’ panels shows *K. pneumoniae* MLSTs, with those  
324 occurring once labelled ‘other’. The final two panels show the Galileo AMR and the TETyper outputs  
325 for the 8 FPs, respectively. The FPs are numbered in ascending order according to abundance in the  
326 hybrid assemblies.

327

328 **Impact statement**

329 The global dissemination of antimicrobial resistance genes (ARGs) has in part been driven by carriage  
330 on mobile genetic elements (MGEs) such as transposons and plasmids. However, our understanding  
331 of these MGEs remains poor, partly due to their high diversity. This means current referenced based  
332 approaches are often inappropriate. ‘Flanker’ is a fast software tool which overcomes this barrier by  
333 *de novo* clustering of ARG flank diversity by sequence similarity. We demonstrate the utility of  
334 Flanker by associating *bla<sub>OXA-48</sub>* and *bla<sub>KPC-2/3</sub>* flanking sequences with geographic regions and  
335 resistance phenotypes.

336

337 **Bibliography**

338 1. **Lipworth S, Vihta K-D, Chau K, Barker L, George S, et al.** Molecular epidemiology of  
339 Escherichia coli and Klebsiella species bloodstream infections in Oxfordshire (UK) 2008-2018.  
340 *medRxiv*.

341 2. **Vihta K-D, Stoesser N, Llewelyn MJ, Quan TP, Davies T, et al.** Trends over time in  
342 Escherichia coli bloodstream infections, urinary tract infections, and antibiotic susceptibilities in  
343 Oxfordshire, UK, 1998–2016: a study of electronic health records. *Lancet Infect Dis*  
344 2018;18:1138–1149.

345 3. **Buetti N, Atkinson A, Marschall J, Kronenberg A, the Swiss Centre for Antibiotic  
346 Resistance (ANRESIS)**. Incidence of bloodstream infections: a nationwide surveillance of acute  
347 care hospitals in Switzerland 2008–2014. *BMJ Open* 2017;7:e013665.

348 4. **Thanner S, Drissner D, Walsh F**. Antimicrobial Resistance in Agriculture. *MBio*  
349 2016;7:e02227-15.

350 5. **Wyres KL, Holt KE**. Klebsiella pneumoniae as a key trafficker of drug resistance genes from  
351 environmental to clinically important bacteria. *Curr Opin Microbiol* 2018;45:131–139.

352 6. **Collis RM, Burgess SA, Biggs PJ, Midwinter AC, French NP, et al.** Extended-Spectrum Beta-  
353 Lactamase-Producing Enterobacteriaceae in Dairy Farm Environments: A New Zealand  
354 Perspective. *Foodborne Pathog Dis* 2019;16:5–22.

355 7. **Velasova M, Smith RP, Lemma F, Horton RA, Duggett NA, et al.** Detection of extended-  
356 spectrum  $\beta$ -lactam, AmpC and carbapenem resistance in Enterobacteriaceae in beef cattle in  
357 Great Britain in 2015. *J Appl Microbiol* 2019;126:1081–1095.

358 8. **von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, et al.**  
359 Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene  
360 Transfer. *Front Microbiol* 2016;7:173.

361 9. **Passarelli-Araujo H, Palmeiro JK, Moharana KC, Pedrosa-Silva F, Dalla-Costa LM, et al.**  
362 Genomic analysis unveils important aspects of population structure, virulence, and antimicrobial  
363 resistance in *Klebsiella aerogenes*. *FEBS J* 2019;286:3797–3810.

364 10. **Nakamura K, Murase K, Sato MP, Toyoda A, Itoh T, et al.** Differential dynamics and  
365 impacts of prophages and plasmids on the pangenome and virulence factor repertoires of Shiga  
366 toxin-producing *Escherichia coli* O145:H28. *Microb Genom*;6. Epub ahead of print January  
367 2020. DOI: 10.1099/mgen.0.000323.

368 11. **Decano AG, Downing T.** An *Escherichia coli* ST131 pangenome atlas reveals population  
369 structure and evolution across 4,071 isolates. *Sci Rep* 2019;9:17394.

370 12. **Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, et al.** SRST2: Rapid genomic  
371 surveillance for public health and hospital microbiology labs. *Genome Med* 2014;6:90.

372 13. **Seemann T.** *mlst*. Github. <https://github.com/tseemann/mlst> (accessed July 12, 2019).

373 14. **Lam MMC, Wick RR, Wyres KL, Holt KE.** Genomic surveillance framework and global  
374 population structure for *Klebsiella pneumoniae*. *Cold Spring Harbor Laboratory*  
375 2020;2020.12.14.422303.

376 15. **Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O.** ClermonTyping: an  
377 easy-to-use and accurate in silico method for *Escherichia* genus strain phlyotyping. *Microb  
378 Genom*;4. Epub ahead of print July 2018. DOI: 10.1099/mgen.0.000192.

379 16. **Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, et al.** Fast and flexible bacterial  
380 genomic epidemiology with PopPUNK. *Genome Res* 2019;29:304–316.

381 17. **Robertson J, Nash JHE.** MOB-suite: software tools for clustering, reconstruction and typing of  
382 plasmids from draft assemblies. *Microb Genom*;4. Epub ahead of print August 2018. DOI:  
383 10.1099/mgen.0.000206.

384 18. **Acman M, van Dorp L, Santini JM, Balloux F.** Large-scale network analysis captures  
385 biological features of bacterial plasmids. *Nat Commun* 2020;11:2452.

386 19. **Sheppard AE, Stoesser N, German-Mesner I, Vigesana K, Sarah Walker A, et al.** TETyper:  
387 a bioinformatic pipeline for classifying variation and genetic contexts of transposable elements  
388 from short-read whole-genome sequencing data. *Microb Genom*;4. Epub ahead of print  
389 December 2018. DOI: 10.1099/mgen.0.000232.

390 20. **Johansson MHK, Bortolaia V, Tansirichaiya S, Aarestrup FM, Roberts AP, et al.** Detection  
391 of mobile genetic elements associated with antibiotic resistance in *Salmonella enterica* using a  
392 newly developed web tool: MobileElementFinder. *J Antimicrob Chemother* 2021;76:101–109.

393 21. **Wang R, van Dorp L, Shaw LP, Bradley P, Wang Q, et al.** The global distribution and spread  
394 of the mobilized colistin resistance gene mcr-1. *Nat Commun* 2018;9:1179.

395 22. **Ludden C, Raven KE, Jamrozy D, Gouliouris T, Blane B, et al.** One Health Genomic  
396 Surveillance of *Escherichia coli* Demonstrates Distinct Lineages and Mobile Genetic Elements in  
397 Isolates from Humans versus Livestock. *MBio*;10. Epub ahead of print January 22, 2019. DOI:  
398 10.1128/mBio.02693-18.

399 23. **David S, Cohen V, Reuter S, Sheppard AE, Giani T, et al.** Integrated chromosomal and  
400 plasmid sequence analyses reveal diverse modes of carbapenemase gene spread among  
401 *Klebsiella pneumoniae*. *Proc Natl Acad Sci U S A*. Epub ahead of print September 23, 2020.  
402 DOI: 10.1073/pnas.2003407117.

403 24. **Acman M, Wang R, van Dorp L, Shaw LP, Wang Q, et al.** Role of the mobilome in the global  
404 dissemination of the carbapenem resistance gene blaNDM. *Cold Spring Harbor Laboratory*  
405 2021;2021.01.14.426698.

406 25. **Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, et al.** Mash: fast genome  
407 and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132.

408 26. **Hendrickx APA, Landman F, de Haan A, Witteveen S, van Santen-Verheuvel MG, et al.**  
409 blaOXA-48-like genome architecture among carbapenemase-producing *Escherichia coli* and  
410 *Klebsiella pneumoniae* in the Netherlands. *Microb Genom*;7. Epub ahead of print May 2021.  
411 DOI: 10.1099/mgen.0.000512.

412 27. **David S, Reuter S, Harris SR, Glasner C, Feltwell T, et al.** Epidemic of carbapenem-resistant  
413 *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. *Nat Microbiol* 2019;4:1919–  
414 1929.

415 28. **Hagberg A, Swart P, S Chult D.** *Exploring network structure, dynamics, and function using*  
416 *NetworkX*. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).  
417 <https://www.osti.gov/biblio/960616> (2008).

418 29. **Wick R.** *Assembly-Dereplicator*. Github. <https://github.com/rrwick/Assembly-Dereplicator>  
419 (accessed February 2, 2021).

420 30. **Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, et al.** *vegan*: Community  
421 Ecology Package. <https://CRAN.R-project.org/package=vegan> (2019).

422 31. **Seemann T.** *abricate*. Github. <https://github.com/tseemann/abricate> (accessed July 5, 2019).

423 32. **EUCAST**. European Committee on Antimicrobial Susceptibility Testing.  
424 [https://www.eucast.org/clinical\\_breakpoints/](https://www.eucast.org/clinical_breakpoints/).

425 33. **Wickham H.** *ggplot2*: Elegant Graphics for Data Analysis. <https://ggplot2.tidyverse.org> (2016).

426 34. **Wilkins D.** *gggenes*: Draw Gene Arrow Maps in “*ggplot2*.” <https://CRAN.R-project.org/package=gggenes> (2019).

428 35. **Yu G, Smith DK, Zhu H, Guan Y, Lam TT.** *Ggtree* : An r package for visualization and  
429 annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol*  
430 *Evol* 2017;8:28–36.

431 36. **Seemann T.** Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.

432 37. **Katz L, Griswold T, Morrison S, Caravas J, Zhang S, et al.** Mashtree: a rapid comparison of  
433 whole genome sequence files. *J Open Source Softw* 2019;4:1762.

434 38. **Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, et al.** In silico  
435 detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing.  
436 *Antimicrob Agents Chemother* 2014;58:3895–3903.

437 39. **Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al.** Biopython: freely available  
438 Python tools for computational molecular biology and bioinformatics. *Bioinformatics*  
439 2009;25:1422–1423.

440 40. **Pitout JDD, Peirano G, Kock MM, Strydom K-A, Matsumura Y.** The Global Ascendancy of  
441 OXA-48-Type Carbapenemases. *Clin Microbiol Rev*;33. Epub ahead of print December 18,  
442 2019. DOI: 10.1128/CMR.00102-19.

443 41. **Beyrouthy R, Robin F, Delmas J, Gibold L, Dalmasso G, et al.** IS1R-mediated plasticity of  
444 IncL/M plasmids leads to the insertion of bla OXA-48 into the Escherichia coli Chromosome.  
445 *Antimicrob Agents Chemother* 2014;58:3785–3790.

446 42. **Carrer A, Poirel L, Eraksoy H, Cagatay AA, Badur S, et al.** Spread of OXA-48-positive  
447 carbapenem-resistant Klebsiella pneumoniae isolates in Istanbul, Turkey. *Antimicrob Agents  
448 Chemother* 2008;52:2950–2954.

449 43. **Chen L, Mathema B, Chavda KD, DeLeo FR, Bonomo RA, et al.** Carbapenemase-producing  
450 Klebsiella pneumoniae: molecular and genetic decoding. *Trends Microbiol* 2014;22:686–696.

451 44. **Cuzon G, Naas T, Nordmann P.** Functional characterization of Tn4401, a Tn3-based  
452 transposon involved in blaKPC gene mobilization. *Antimicrob Agents Chemother* 2011;55:5370–  
453 5373.

454





