Flanker: a tool for comparative genomics of gene flanking regions

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

* Joint first authors

^ Joint senior authors

[1] Nuffield Department of Medicine, University of Oxford, UK

[2] Oxford University Hospitals NHS Foundation Trust, Oxford, UK

[3] NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, United Kingdom

[4] NIHR Oxford Biomedical Research Centre, Oxford, UK

[5] National Infection Service, Public Health England, Colindale, London, UK

[6] Department of Zoology, University of Oxford, UK

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

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

Abstract

Analysing the flanking sequences surrounding genes of interest is often highly relevant to understanding the role of mobile genetic elements (MGEs) in horizontal gene transfer, particular for antimicrobial resistance genes. Here, we present Flanker, a Python package which performs alignment-free clustering of gene flanking sequences in a consistent format, allowing investigation of MGEs without prior knowledge of their structure. Flanker clusters flanking sequences based on Mash distances, allowing for easy comparison of similarity and the extent of this similarity across sequences. Additionally, Flanker can be flexibly parameterised to finetune outputs by characterising upstream and downstream regions separately and investigating variable lengths of flanking sequence. We apply Flanker to two recent datasets describing plasmid-associated carriage of important carbapenemase genes (*bla*$_{\text{OXA-48}}$ and *bla*$_{\text{KPC-2/3}}$) and show that it successfully identifies distinct clusters of flanking sequences (flank patterns), including both known and previously uncharacterised structural variants. We demonstrate that flank patterns are linked to geographical regions and carbapenem phenotypes, suggesting they may be useful as epidemiological markers. Flanker is freely available under an MIT license at https://github.com/wtmatlock/flanker.

Data Summary

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

Introduction

The increasing incidence antimicrobial resistance (AMR) in clinical isolates poses a threat to all areas of medicine [1–3]. AMR genes (ARGs) are found in a diverse range of genetic contexts, bacterial species, and in both clinical and non-clinical environments (e.g., agricultural, refuse and natural ecosystems) [4–7]. However, the mechanisms underpinning the dissemination of many ARGs

between these reservoirs remain poorly understood, limiting the efficacy of surveillance and the ability to design effective interventions. Usually, ARGs are spread vertically (either via chromosomal integration or stable association of a plasmid within a clonal lineage) or by horizontal gene transfer (HGT) through mobile genetic elements (MGEs) e.g., transposons or plasmids. HGT can accelerate the rate of ARG acquisition, both within and across species.

The epidemiology of ARGs can therefore involve multiple levels, from clonal spread to MGEs. There are many existing software tools to facilitate epidemiological study of bacterial strains [8–12], whole plasmids [13, 14], and smaller MGEs[15, 16]. Several tools and databases exist for the annotation of non-plasmid MGEs such as insertion sequences (ISs) and transposons [15, 16], but all rely on comparisons to reference sequences, so are limited to known diversity. Reference-free tools for analysing MGE diversity would therefore be a useful addition. Here we describe a Flanker, a simple, reference-free tool to investigate MGEs by analysing the flanking sequences of ARGs.

The flanking sequences (hereafter, ‘flanks’) around an ARG that has been mobilised horizontally may act as signatures of relevant MGEs and support epidemiological analyses. However, these flanks can contain a great deal of structural variation due to their evolutionary history. Where a single known MGE is under investigation, it is possible to specifically type this element (for example, using TETyper [15]) or align flanks against a known ancestral form after the removal of later structural variation [17]. However, often multiple structures may be involved. This is particularly true for ARGs which move frequently on a variety of MGEs. Studies of different ARGs often choose different *ad hoc* approaches to extract flanks and cluster genetic structures. Examples include hierarchical clustering of isolates carrying an ARG based on short-read coverage of known ARG-carrying contigs [18], assigning assembled contigs into 'clustering groups' based on gene presence and synteny [19] or iterative 'splitting' of flanks based on pairwise BLAST [20]. A consistent and simple approach for this task would not only avoid repeated method development, but also aid comparison between methods developed for specific ARGs.

To address this problem, we developed Flanker, a pipeline to analyse the regions around a given ARG in a consistent manner. Flanker flexibly extracts the flanks of a specified gene from a dataset of

contigs, then clusters these sequences using Mash distances to identify consistent structures [21]. Flanker is available as a documented Python and Bioconda package released under the MIT open-source license. Source code is deposited at https://github.com/wtmatlock/flanker and documentation at https://flanker.readthedocs.io/en/latest/.

Methods

Datasets

To validate Flanker, demonstrate its application and provide a comparison with existing tools, we used two recent datasets of complete plasmids (derived from hybrid long-/short-read assemblies) containing carbapenemase genes of clinical importance [19, 24]. The first dataset comprised 51 $bla_{\text{OXA-48}}$-harbouring plasmids in carbapenem-resistant *E. coli* and *K. pneumoniae* isolates from patients in the Netherlands (2014-2019) [26]; the second dataset comprised 50 $bla_{\text{KPC-2}}$ or $bla_{\text{KPC-3}}$-harbouring plasmids in carbapenem-resistant *K. pneumoniae* isolated as part of the EuSCAPE study [23], a European surveillance programme investigating carbapenem resistance in Enterobacterales. The EuSCAPE dataset [19, 25] additionally contains a large collection of short-read sequencing data for *Klebsiella* spp. isolates alongside meropenem susceptibility data. This was used to demonstrate additional possible epidemiological applications of the Flanker tool by evaluating whether specific flank patterns were more likely to be associated with phenotypic meropenem resistance.

Mash distances

Pair-wise distances between flanks were calculated using Mash (version 2.2.2) [21]. Mash reduces sequences to a fixed-length MinHash sketch, which is used to estimate the Jaccard distance between k-mer content. It also gives the Mash distance, which ranges from 0 (~identical sequences) to 1 (~completely dissimilar sequences). We used the default Mash parameters in all analyses. The Mash distance was developed to approximate the rate of sequence mutation between genomes under a simple evolutionary model, and explicitly does not model more complex processes. We use it here for

fast alignment-free clustering of sequences and do not draw any direct conclusions about evolution from pairwise comparisons.

Clustering

To cluster the flanks, Flanker generates an adjacency matrix weighted by Mash distances. It then thresholds this matrix to retain edges weighted less than or equal to the defined threshold. This is then used to construct a graph using the Python NetworkX library [26] and clusters are defined using the nx.connected_components function, which is analogous to single linkage. This is a similar methodology to that used by the Assembly De-replicator tool [27] (from which Flanker re-uses several functions). However, Flanker aims to assign all flanks to a cluster rather than to deduplicate by cluster.

Cluster validation

We validated the output of flanking sequence-based clustering was validated using a PERMANOVA test, implemented with the adonis function from the Vegan package (version 2.7.5) [28] in R. Only flanks in clusters of at least two members were considered (42/51 and 48/50 of $bla_{\text{OXA-48}}$ and $bla_{\text{KPC-2/3}}$ flanks, respectively). The formula used was Mash dist ~ cluster, with the 'euclidean' method and 999 permutations.

Comparison to existing methods/application

We compared the classifications of TETyper (v1.1) [15] and MEFinder (v1.0.3) [16] to those produced by Flanker for 500bp and 5000bp flanks around $bla_{\text{KPC-2/3}}$ genes. TETyper was run using the –threads 8 and --assemblies options with the Tn*4401* reference and SNP/structural profiles provided in the package and MEFinder was run in Abricate using the –mincov 10 option. For comparisons of the proportions of resistant isolates per flank pattern (denoted FPx), isolates were classified as resistant or sensitive using the European Committee on Antimicrobial Susceptibility Testing (EUCAST) breakpoint for meropenem (>8mg/L) [29].

Data visualisation

All figures were made in using Biorender (https://biorender.com) and the R packages ggplot2 (v3.3.0) [30], gggenes (v0.4.0) [31] and ggtree (v2.4.1) [32] Prokka (v1.14.6) [33] was used to annotate Flanker output. Mashtree (v1.2.0) [34] was used to construct a visual representation of Mash distances between whole plasmid genomes. Plasmidfinder was used to detect the presence/absence of plasmid types using Abricate (version 1.0

1) with --mincov 80 and --minid 80 [35].

Results

Flanker

The Flanker package contains two basic modules: the first extracts a region of length *w* around an annotated gene of interest, and the second clusters such regions based on a user-defined Mash distance threshold (default --threshold 0.001, Fig. 1a). Within each FASTA/multi-FASTA format input file, the location of the gene of interest is first determined using the Abricate annotation tool [22]. Flanks around the gene (optionally including the gene itself to enable complete alignments with --include_gene) are then extracted and written to a FASTA format file using BioPython [23]. Flanker gives users the option to either extract flanks using a single window (defined by length in base-pairs [bp]) or multiple windows from a start position (--window) to an end position (--wstop) in fixed increments (--wstep). Flanks may be extracted from upstream, downstream or on both sides of the gene of interest (--flanks). Corrections are also made for circularised genomes where the gene occurs close to the beginning or end of the sequence (--cluster mode) and for genes found on both positive and negative strands. The clustering module groups flanks of user-defined sequence lengths together based on a user-defined Mash [21] distance threshold (--threshold) of user-defined sequence lengths.

In default mode (--mode default), Flanker considers multiple gene queries in turn. In multi-allelic mode (--mode mm), Flanker considers all genes in the list for each window (Fig. 1b). Multiple genes can be queried by either a space-delimited list in the command line (--gene geneA geneB), or a newline-delimited file with the list of genes option (--list_of_genes). A supplementary module ‘salami

mode' (--mode sm) is provided to allow comparison of non-contiguous blocks from a start point (--window), step size (--wstep) and end point (--wstop) (Fig. 1c).

Clustering validation and comparison with TETyper/MEFinder

The clustering mode was validated numerically with a PERMANOVA test (Mash dist ~ cluster: *bla*$_{\text{OXA-48}}$ *p*-value < 0.001, *bla*$_{\text{KPC2/3}}$ *p*-value < 0.001; see Methods). Figures 2 and 3 also provide a visual comparison of an alignment of genes ('Alignment' panel) to the cluster labels (i.e., flank pattern).

Of the two existing tools we compared in evaluating the flanks around *bla*$_{\text{KPC2/3}}$, TETyper was by far the slowest (1172 seconds [s]), whereas MEFinder (run in Abricate) and Flanker took 7s and 11s, respectively (benchmarked on 5000bp upstream flanks on a cluster with Intel Skylake 2.6GHz chips). MEFinder was able to detect Tn*4401* but could not provide any further structural resolution and was unable to classify 6/50 (12%) 500bp and 1/50 (2%) 5000bp flanks. TETyper structural profiles were consistent with Flanker when analysing 500 and 5000bp upstream regions, though Flanker split a group of six isolates with the TETyper structural profile 1-7127|7202-10006 into four groups (Table S2). To map our FPs to the established nomenclature, we additionally compared the output of Flanker to that of TETyper when the latter was given the entire Tn*4401* sequence (i.e., by evaluating the typical 7,200bp Tn*4401*-associated flank upstream of blaKPC). Flanker and TETyper classifications of Tn*4401* regions were broadly consistent (Table S2), though this analysis demonstrated the potential benefit of the reference-free approach of Flanker which showed that four non-Tn*4401* structural profiles ('unknown' in TETyper) were distinct from each other. In addition, TETyper classified 3 flanks as Tn*4401*_truncC-1, whereas Flanker resolved this cluster into two distinct groups (Table S2).

Application to plasmids carrying *bla*$_{\text{OXA-48}}$

The carbapenemase gene *bla*$_{\text{OXA-48}}$ has been shown to be disseminated by Tn*1999*-associated structures (~5kb) nested in L/M-type plasmids, and as part of an IS1R-associated composite transposon containing *bla*$_{\text{OXA-48}}$ and part of Tn*1999*, namely Tn*6237* (~21.9kb), that has been

implicated in the chromosomal integration of *bla*$_{OXA-48}$ [25, 36]. It has been recently demonstrated that most *bla*$_{OXA-48}$-like genes in clinical isolates in Europe are carried on highly similar L/M(pOXA-48)-type plasmids, with evidence of both horizontal and vertical transmission across a diverse set of sequence types [19]. Whilst Tn*1999*-like flanking regions are relatively well characterised [37], in this example we chose an initial arbitrary upstream window of 5000bp to simulate a scenario in which there is no prior knowledge. Inspection of a plot of window clusters (i.e., as shown in the 'Flankergram' in Fig. 2) demonstrates that Flanker output allows the empirical identification of the position ~2200bp upstream of *bla*$_{OXA-48}$ as an important point of structural divergence without requiring annotation (as shown at ~2200 along the *x*-axis, where the window cluster colour schemes diverge), corresponding to the edge of Tn*1999* at its expected position.

Flanker grouped flanks into seven distinct flank patterns (FPs) in the 5000bp sequence upstream of *bla*$_{OXA-48}$ in isolates carrying L/M(pOXA-48)-type plasmids (Fig. 2). To investigate the association of phenotypic carbapenem resistance with *bla*$_{OXA-48}$ FPs, we created a Mash sketch using one randomly chosen representative per group and screened an Illumina sequenced collection of European carbapenemase-resistant *Klebsiella* isolates[25] for containment (n=425). Mash screen, assigning FP based on the top hit [median identity = 1.00; range: 0.93-1.00]). Two FPs (FP3 and FP7) accounted for 394/425 (93%) of isolates; both were widely distributed across Europe. Of the 226 isolates with meropenem susceptibility data available, those belonging to FP7 were proportionally more meropenem-resistant compared to FP3 (72/142 (51%) vs. 16/73 (22%), exact *p*-value<0.001 (Fig S1). Annotation (using Gallileo AMR; see methods) of these revealed that whereas FP3 contains Tn*1999*, FP7 contains Tn*1999.2*, which has been previously described as creating a strong promoter which produces 2-fold higher enzymatic activity [38].

<u>Application to plasmids carrying *bla*$_{KPC-2/3}$</u>

David et al. showed that *bla*$_{KPC-2/3}$ genes have been disseminated in European *Klebsiella pneumoniae* clinical isolates via a diverse collection of plasmids in association with a dominant clonal lineage, ST258/512, which accounted for 230/312 (74%) of *bla*$_{KPC}$-associated isolates in the EuSCAPE collection [19]. *bla*$_{KPC}$ has largely been associated with variants of a ~10kb transposon, Tn*4401*[38,

39]. From the collection of 50 hybrid assembled KPC-containing plasmids, Flanker grouped 8 distinct FPs over a 7200bp window upstream of $bla_{KPC-2/3}$ (Fig. 3). This window length was chosen to capture the entire Tn*4401* sequence upstream of bla_{KPC}. Of the FPs, 346/442 (78%) belonged to FP1 (corresponding to isoform Tn*4401*a). Whilst FP1 was widely distributed across Europe, FPs 2 (corresponding to Tn*4401*_truncC) and FP7 (corresponding to Tn*4401*d) appeared more geographically restricted: FP2 to Slovakia (5/5, 100%) and FP7 to Israel (19/59, 32%) and Portugal (34/59, 58%) with isolates also found in Poland and Germany (n=2 each) and Italy and Austria (n=1). There was no evidence of a difference in the proportion of isolates resistant to meropenem between FP1 and FP7 (202/238 (85%) vs 23/25 (92%), exact *p*-value=0.5), though there was incomplete susceptibility data for isolates from both groups (108/346 [31%] for FP1 and 38/63 [60%] for FP7).

Discussion

We present Flanker, a fast and flexible Python package for analysing gene flanking sequences. We anticipate that this kind of analysis will become more common as the number of complete reference-grade, bacterial assemblies increase. Our analysis of data from the EuSCAPE project demonstrated Flanker's ability to reproduce existing knowledge and suggests that flank patterns (FPs) might additionally be useful epidemiological markers, for example in evaluating geographical associations of sequences or associations with resistance phenotypes.

A key advantage compared to existing tools is that there is no reliance on reference sequences or prior knowledge. Despite analysing only a relatively small number (n=50) of complete bla_{KPC} containing plasmids, there were four distinct FPs which TETyper classified as 'unknown' because their profiles were not in previously known structural profiles. TETyper works well when alleles/structural variants are known but can only classify a single transposon type at a time and requires manual curation when this is not the case. The observed diversity of flanking sequences is likely to continue to increase and manual curation of naming schemes will be arduous to maintain. MEFinder on the other hand is a quick screening tool which can search a large library of known mobile elements but lacks sequence level resolution. Whilst Flanker overcomes these challenges, users may need to perform downstream analysis to interpret its output. We hope that Flanker will be complementary to these and other similar

existing tools by reducing the dimensionality of large datasets and identifying smaller groups of sequences to focus on in detail. Though we have developed Flanker for ARGs, Abricate allows use of custom databases meaning any desired genes of interest could be analysed. Accurate outputs from Flanker will be dependent on the quality of input assemblies, and on the correct annotation of the gene of interest.

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

Authors and contributors

Contributions have been attributed by the CRediT system as follows:

Conceptualisation: WM, SL, LPS, NS

Methodology: WM, SL

Software: WM, SL, BC

Validation: WM, SL

Formal Analysis: WM, SL

Investigation: WM, SL

Resources: DWC, TP, ASW, NS

Data Curation: SL, WM

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

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

Visualisation: SL, WM

Supervision: LPS, NS, TP, ASW, DC

Project Administration: SL, WM, NS, LPS

Funding: TP, DC, ASW, NS

Data availability

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

Conflicts of Interest

The authors have no conflicts of interest.

Funding Information

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

Acknowledgements

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

Figure Descriptions

Figure 1: Schematic of Flanker's modes and parameters. (a) Flanker uses Abricate to annotate the gene of interest in input sequences and outputs associated flanking sequences, optionally clustering these on a user defined Mash distance threshold. It can take linear or circularised sequences. (b) In

this example, genes *bla*$_{KPC-2}$ and *bla*$_{KPC-3}$ have been queried (-g blaKPC-2 blaKPC-3), and only the upstream flank is desired (-f upstream). The top single black arrow represents choosing a single window of length 3000bp (-w 3000), whereas the bottom three black arrows represent stepping in 1000bp windows from 0bp to 3000bp (-w 0 –wstep 1000 -wstop 3000). The default mode (-m default) processes all annotated alleles separately, but the multi-allelic mode (-m mm) processes all alleles together. (c) Flanker has a supplementary salami mode (-m sm), which outputs non-contiguous blocks of sequence with a start point, step size, and end point (-w 0 -wstep 1000 -wstop 3000), represented by the three black arrows.

Figure 2: Flanking regions 5000bp upstream of *bla*$_{OXA-48}$ carrying plasmids from *Klebsiella pneumoniae* isolates. The 'Tree' panel is a Neighbour-Joining tree constructed from Mash distances between complete sequences of plasmids carrying the *bla*$_{OXA-48}$ gene. The second panel indicates the presence/absence of a L/M(pOXA-48)-type plasmid. The 'Alignment' panel schematically represents coding regions in the 5000bp sequence upstream of the *bla*$_{OXA-48}$ gene, which is shown in red. Other genes are coloured according to the flank pattern which considers the overall pattern of all 100bp window clusters up to 2200bp (the approximate upstream limit of Tn*1999*). The "Flankergram" shows window clusters of all groups over each 100bp window between 0 and 5000bp. The 'MLST' panel shows *K. pneumoniae* multi-locus sequence types (MLSTs), with those occurring once labelled 'other'.

Figure 3: Flanking regions 7200bp upstream of *bla*$_{KPC-2/3}$ carrying plasmids from *Klebsiella pneumoniae* isolates. The 'Tree' panel is a Neighbour-Joining tree constructed from Mash distances between complete sequences of plasmids carrying the *bla*$_{KPC-2/3}$ gene. The next three panels indicate the presence/absence of FIB(pQ1I)-, FII(pKP91)-, and FIB(Kpn3)-type plasmids. The 'Gene' column indicates which *bla*$_{KPC}$ allele (2 or 3) is present. The 'Alignment' panel schematically represents coding regions in the 7200bp sequence region upstream of the *bla*$_{KPC-2/3}$ gene, which is shown in red. Other genes are coloured according to the flank group, which here takes into account the overall pattern of all 100bp window groups (shown in the "Flankergram" panel) over the full 7200bp region upstream of *bla*$_{KPC-2/3}$. The "Flankergram" shows window clusters over each 100bp window between

0 and 7200bp. The ‘MLST’ panels shows *K. pneumoniae* MLSTs, with those occurring once labelled ‘other’.

Bibliography

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

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

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

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

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

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

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

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

9. **Seemann T**. *mlst*. Github. https://github.com/tseemann/mlst (accessed July 12, 2019).

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

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

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

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

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

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

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

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

18. **Ludden C, Raven KE, Jamrozy D, Gouliouris T, Blane B, *et al.*** One Health Genomic Surveillance of Escherichia coli Demonstrates Distinct Lineages and Mobile Genetic Elements in

Isolates from Humans versus Livestock. *MBio*;10. Epub ahead of print January 22, 2019. DOI: 10.1128/mBio.02693-18.

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

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

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

22. **Seemann T**. *abricate*. Github. https://github.com/tseemann/abricate (accessed July 5, 2019).

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

24. **Hendrickx APA, Landman F, de Haan A, Borst D, Witteveen S, *et al.*** blaOXA-48-like genome architecture among carbapenemase-producing Escherichia coli and Klebsiella pneumoniae in the Netherlands. *Cold Spring Harbor Laboratory* 2020;2020.12.18.423568.

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

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

27. **Wick R**. *Assembly-Dereplicator*. Github. https://github.com/rrwick/Assembly-Dereplicator (accessed February 2, 2021).

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

29. **EUCAST**. European Committee on Antimicrobial Susceptibility Testing. https://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_10.0_Breakpoint_Tables.pdf.

30. **Wickham H**. ggplot2: Elegant Graphics for Data Analysis. https://ggplot2.tidyverse.org (2016).

31. **Wilkins D**. gggenes: Draw Gene Arrow Maps in "ggplot2." https://CRAN.R-project.org/package=gggenes (2019).

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

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

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

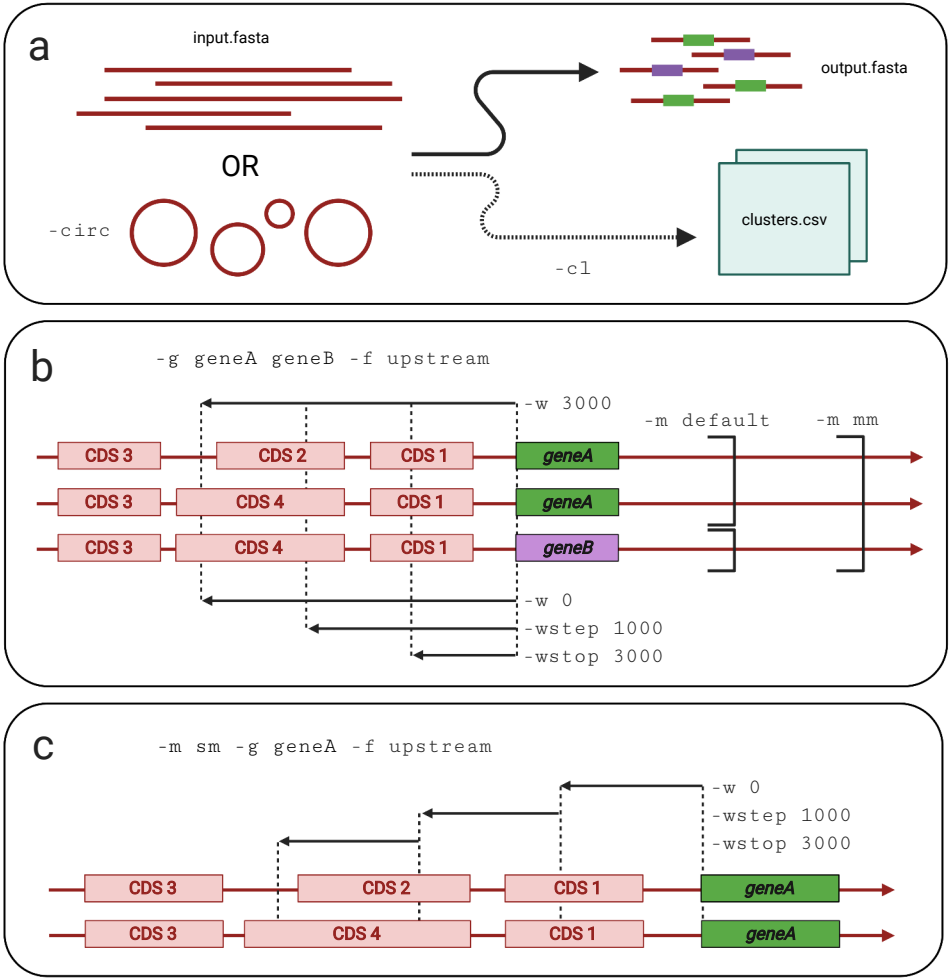
35. **Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, *et al.*** In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58:3895–3903.

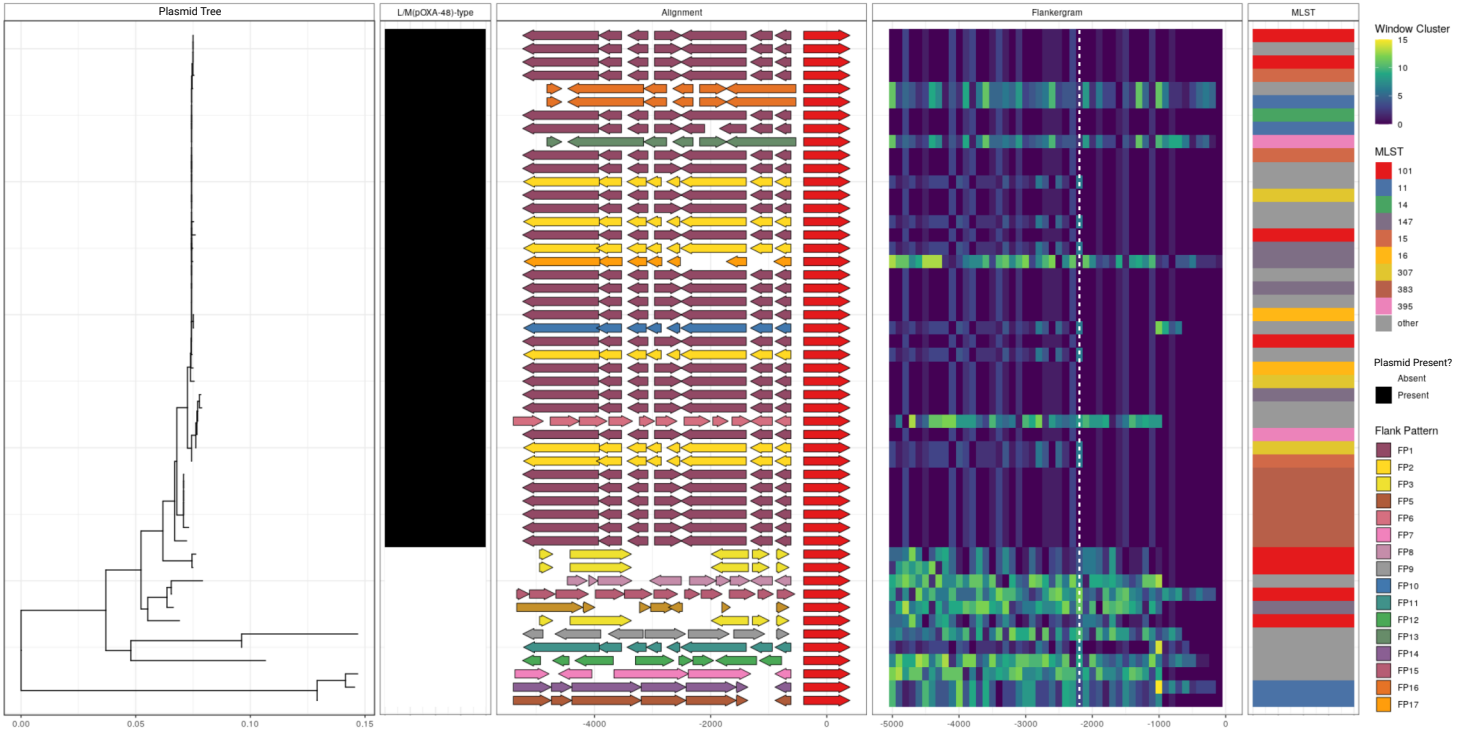
36. **Beyrouthy R, Robin F, Delmas J, Gibold L, Dalmasso G, *et al.*** IS1R-mediated plasticity of IncL/M plasmids leads to the insertion of bla OXA-48 into the Escherichia coli Chromosome. *Antimicrob Agents Chemother* 2014;58:3785–3790.

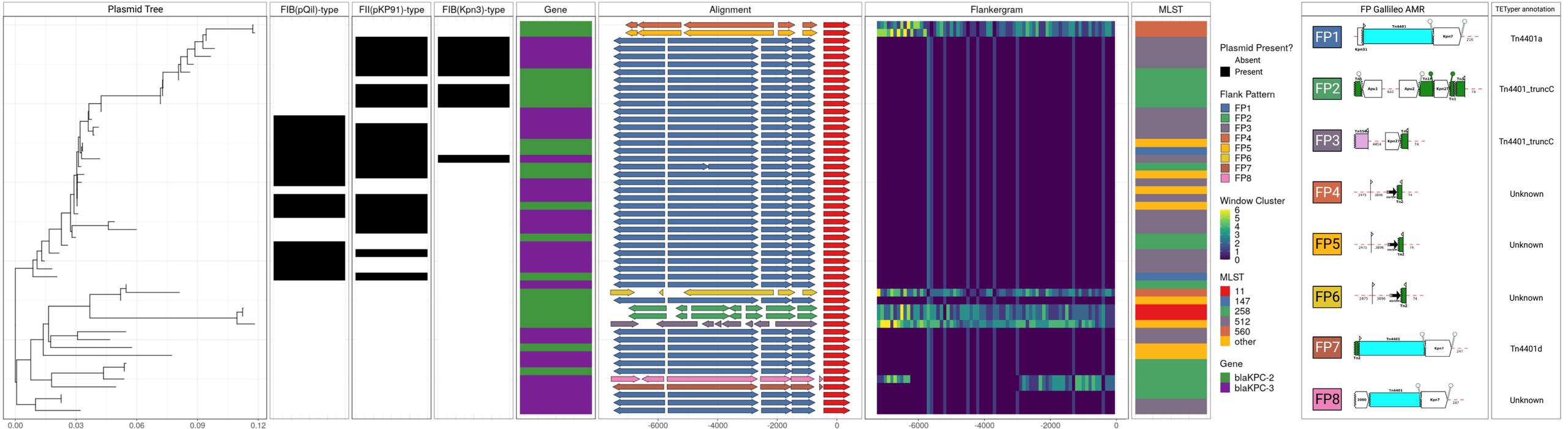
37. **Pitout JDD, Peirano G, Kock MM, Strydom K-A, Matsumura Y**. The Global Ascendency of OXA-48-Type Carbapenemases. *Clin Microbiol Rev*;33. Epub ahead of print December 18, 2019. DOI: 10.1128/CMR.00102-19.

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

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


a
input.fasta
OR
-circ
output.fasta
clusters.csv
-cl
b
-g geneA geneB -f upstream
-w 3000
-m default
-m mm
CDS 3
CDS 2
CDS 1
geneA
CDS 3
CDS 4
CDS 1
geneA
CDS 3
CDS 4
CDS 1
geneB
-w 0
-wstep 1000
-wstop 3000
c
-m sm -g geneA -f upstream
-w 0
-wstep 1000
-wstop 3000
CDS 3
CDS 2
CDS 1
geneA
CDS 3
CDS 4
CDS 1
geneA


Plasmid Tree
L/M(pOXA-48)-type
Alignment
Flankergram
MLST
0.00
0.05
0.10
0.15
-4000
-2000
0
-5000
-4000
-3000
-2000
-1000
0
Window Cluster
15
10
5
0
MLST
101
11
14
147
15
16
307
383
395
other
Plasmid Present?
Absent
Present
Flank Pattern
FP1
FP2
FP3
FP5
FP6
FP7
FP8
FP9
FP10
FP11
FP12
FP13
FP14
FP15
FP16
FP17


Plasmid Tree
FIB(pQil)-type
FII(pKP91)-type
FIB(Kpn3)-type
Gene
Alignment
Flankergram
MLST
FP Gallileo AMR
TETyper annotation
0.00
0.03
0.06
0.09
0.12
-6000
-4000
-2000
0
Plasmid Present?
Absent
Present
Flank Pattern
FP1
FP2
FP3
FP4
FP5
FP6
FP7
FP8
Window Cluster
6
5
4
3
2
1
0
MLST
11
147
258
512
560
other
Gene
blaKPC-2
blaKPC-3
FP1
FP2
FP3
FP4
FP5
FP6
FP7
FP8
Tn4401a
Tn4401_truncC
Tn4401_truncC
Unknown
Unknown
Unknown
Tn4401d
Unknown