MICROBIAL GENOMICS

RESEARCH ARTICLE

Shaw and Neher, *Microbial Genomics* 2023;9:001168

DOI 10.1099/mgen.0.001168

Visualizing and quantifying structural diversity around mobile resistance genes

Liam P. Shaw[1,2,*] and Richard A. Neher[3]

Abstract

Understanding the evolution of mobile genes is important for understanding the spread of antimicrobial resistance (AMR). Many clinically important AMR genes have been mobilized by mobile genetic elements (MGEs) on the kilobase scale, such as integrons and transposons, which can integrate into both chromosomes and plasmids and lead to rapid spread of the gene through bacterial populations. Looking at the flanking regions of these mobile genes in diverse genomes can highlight common structures and reveal patterns of MGE spread. However, historically this has been a largely descriptive process, relying on gene annotation and expert knowledge. Here we describe a general method to visualize and quantify the structural diversity around genes using pangraph to find blocks of homologous sequence. We apply this method to a set of 12 clinically important beta-lactamase genes and provide interactive visualizations of their flanking regions at https://liampshaw.github.io/flanking-regions. We show that nucleotide-level variation in the mobile gene itself generally correlates with increased structural diversity in its flanking regions, demonstrating a relationship between rates of mutational evolution and rates of structural evolution, and find a bias for greater structural diversity upstream. Our framework is a starting point to investigate general rules that apply to the horizontal spread of new genes through bacterial populations.

DATA SUMMARY

Analysis pipeline (release: v1.0.0): https://github.com/liampshaw/mobile-gene-regions. Stably archived at https://doi.org/10.5281/zenodo.10213206.

Interactive plots for n=12 beta-lactamases:

https://liampshaw.github.io/flanking-regions.

Beta-lactamase dataset:

https://doi.org/10.5281/zenodo.8208376.

INTRODUCTION

Mobile genetic elements (MGEs) allow the horizontal movement of genes within and between bacterial species, contributing to a vast range of bacterial phenotypes, including antimicrobial resistance (AMR). A 'mobile gene' will be found in multiple genomic contexts, where its flanking regions contain information on the evolutionary history of the MGE (or MGEs) that have mobilized it [1]. However,

[1] This suggests that bla$_{CTX-M-14}$ is the ancestral gene, so would be a better choice of focal gene than blaC$_{TX-M-65}$ as we arbitrarily chose here. In fact CTX-M-14 has been suggested to have evolved twice by convergent evolution with different flanking regions for the two nucleotide variants [31].

Received 29 September 2023; Accepted 07 December 2023; Published 20 December 2023

Author affiliations: [1]Department of Biology, University of Oxford, Oxford, UK; [2]Department of Biosciences, University of Durham, Durham, UK; [3]Biozentrum, University of Basel, Basel, Switzerland.

*Correspondence: Liam P. Shaw, liam.shaw@biology.ox.ac.uk

Keywords: AMR; beta-lactamases; mobile genetic elements; *Enterobacterales*; microbial evolution; plasmids.

Abbreviations: AMR, antimicrobial resistance; ESBL, extended-spectrum beta-lactamase; IS, insertion sequence; MGE, mobile genetic element; SNV, single nucleotide variant.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.Two supplementary figures and one supplementary table are available with the online version of this article.

Impact Statement

Understanding the evolution and spread of mobile resistance genes is challenging because of the high variability in their genomic contexts. Here we outline a fast computational approach that identifies stretches of homologous sequence in the flanking regions of a gene, simultaneously producing interactive visualizations of these regions and quantifying the diversity within them. As an example, we apply the method to 12 clinically important beta-lactamase genes. We find that structural diversity around the resistance gene is correlated with mutations within it and that there is greater structural diversity upstream of many genes. There may be other general patterns about the evolution of mobile resistance genes that can be recovered with this kind of analysis.

despite the increasing availability of complete genomes, the high structural diversity in these flanking regions generated by both transposition and recombination means that analysing them is challenging.

Each mobile AMR gene has its own unique epidemiological history. As a recent example, the metallobeta-lactamase gene $bla_{\mathrm{NDM-1}}$ was first reported by Yong *et al*. [1]. The earliest NDM-positive isolate dates only from 2005 [2], but already by the 2010s $bla_{\mathrm{NDM-1}}$ had been seen worldwide in diverse bacteria. As of 2023 there are public genomes containing bla_{NDM} genes from 17 bacterial genera [3]. Such a rapid horizontal and global spread with multiple rearrangements presents a challenge for genomic epidemiology. The flanking regions around $bla_{\mathrm{NDM-1}}$ show large structural diversity, particularly upstream of the gene, although with some traces of a common ancestral MGE. Acman *et al*. [4] developed a methodology for iteratively 'splitting' flanking sequences, finding that downstream patterns of structural diversity in a global dataset supported the previous conclusion that $bla_{\mathrm{NDM-1}}$ had been first mobilized by a Tn*125* transposon [5]. This example demonstrates that analysis of flanking regions is a valuable but difficult task, often requiring bespoke methods. For this reason, almost all the existing literature on the flanking regions of mobile genes remains descriptive and focused on single genes at a time, meaning that it is difficult to extrapolate to the general rules – if any – that govern evolution in these regions.

The high levels of structural diversity around mobile genes present two challenges: visualization and quantification. Visualizations are often based on annotated gene clusters [6], but genes are frequently disrupted in flanking regions. Other tools aim to identify discrete clusters for tracking MGE epidemiology. For example, TETyper was developed specifically for transposable MGEs and can identify small-scale changes associated with transposition [7] and Flanker performs alignment-free clustering of flanking sequences based on mash distances [8]. Such tools are useful but difficult to connect to the processes that generate structural diversity. In general, our understanding of these processes remains qualitative.

Here, we aim to provide a starting point for quantitative analysis of structural diversity around mobile genes. We outline an annotation-free approach based on finding homologous sequence blocks with pangraph [9] that scales to thousands of sequences and connects visualization with quantification. As a demonstration, we apply our method to the flanking regions of 12 beta-lactamase genes.

METHODOLOGY

Overview of the pipeline

Fig. 1a gives a schematic overview of the pipeline available at https://github.com/liampshaw/mobile-generegions. The use case is where an investigator has a focal gene and a set of contigs. For example, the focal gene might be $bla_{\mathrm{NDM-1}}$ and the contigs would be assemblies containing bla_{NDM} gene variants downloaded from National Center for Biotechnology Information (NCBI) MicroBIGG-E with the search query element_symbol:blaNDM*. The first step is to locate the focal gene, using BLASTN (default parameters) followed by retaining only hits matching lateral coverage and nucleotide-level difference thresholds which can be specified by the user [default: >99% coverage, <25 nucleotide-level differences in the alignment, with 1 single-nucleotide variant (SNV) counted as 1 difference and an *n*-bp indel counted as *n* differences; in practice, almost all variants are SNVs so we refer to this threshold as an SNV threshold]. This location is then used to extract the flanking regions for some specified distance value (default: +/− 5 kb upstream and downstream). By default, the pipeline orientates the extracted region so that the focal gene is on the positive strand. It also omits contigs that contain more than one copy of the focal gene and those that are shorter than the requested flanking distance (schematic examples of excluded contigs are shown in Fig. 1a).

The pipeline then uses pangraph [9] to find homologous sequences. The pangraph algorithm was developed for whole genomes, and aims to identify stretches of homologous sequence within and across all input genomes, approximating multiple-genome alignment through iterative pairwise alignment of subsets of sequences with either minimap2 or mmseqs2. These stretches of homologous sequence are referred to as 'pancontigs': linear multiple-sequence alignments, with small indels and nucleotide polymorphisms retained in the underlying data structure. The algorithm aligns pairs of subgraphs in a traversal of a guide tree of

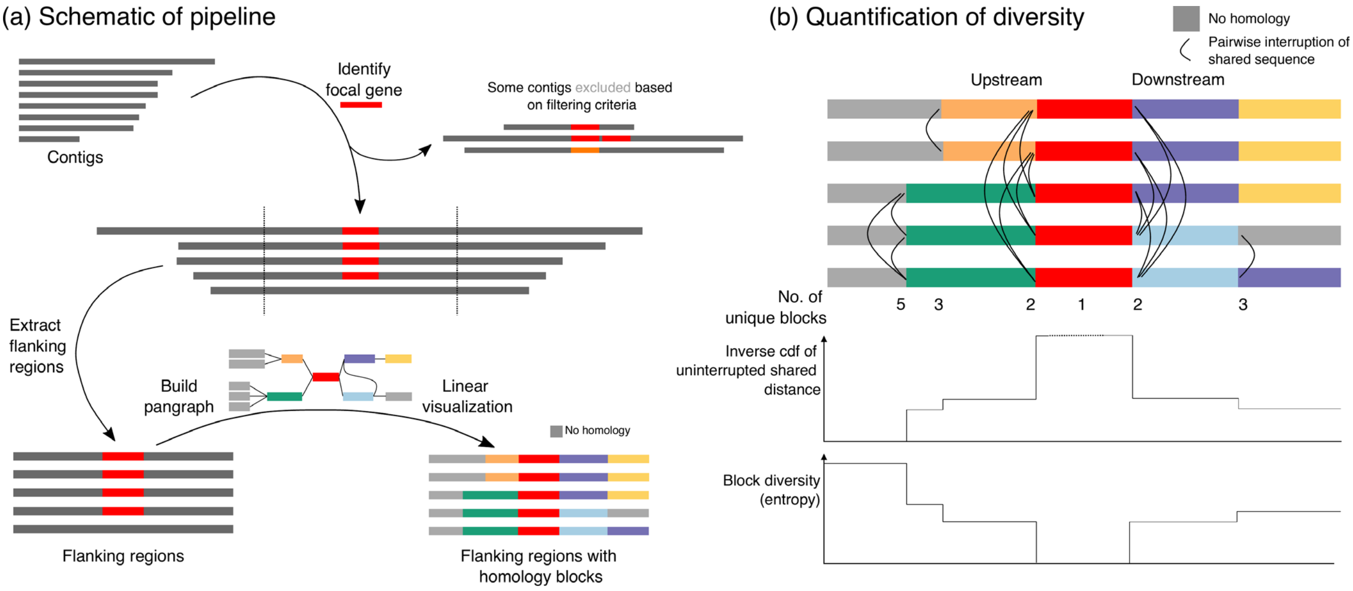
(a) Schematic of pipeline
Contigs
Identify focal gene
Some contigs excluded based on filtering criteria
Extract flanking regions
Build pangraph
Linear visualization
No homology
Flanking regions
Flanking regions with homology blocks
(b) Quantification of diversity
No homology
Pairwise interruption of shared sequence
Upstream
Downstream
No. of unique blocks
5
3
2
1
2
3
Inverse cdf of uninterrupted shared distance
Block diversity (entropy)

Fig. 1. Overview of pipeline and quantification of structural diversity. (a) Given a focal gene and a set of contigs as input, the pipeline extracts the gene's flanking regions (default: 5 kb upstream/downstream). Contigs are filtered based on user-defined criteria, removing those with >1 copy of the gene, excessive diversity in the focal gene (default: >25 nucleotide-level differences), or inadequate lateral coverage of the gene (default: <99%). The extracted flanking regions are then passed into pangraph to identify blocks of homologous sequence and create a graph data structure. This graph data structure can be used to visualize the original flanking regions with blocks coloured by homology in a linear layout, highlighting shared homology across input contigs. Optionally, annotations (in gff format) can also be included. This linear visualization is produced as an interactive html output. Blocks without homology to other blocks in the input contigs are coloured grey (i.e. unique sequence). (b) For any pairwise comparison of flanking regions, we define the 'uninterrupted shared distance' as the upstream/downstream distance to their first breakpoint in homology (i.e. indicated by different coloured blocks or by grey blocks). The inverse cumulative distribution function of these pairwise uninterrupted shared distances shows how structural similarity falls away with distance from the focal gene in both directions. The block diversity, computed as the entropy of the vector of blocks at any linear position, gives another way of measuring structural diversity at distances from the focal gene.

all input sequences, using a score to determine the most favourable mergers of aligned pancontigs, which takes into account the alignment length, the creation of additional pancontigs and the number of polymorphisms. The output of pangraph summarizes a set of input genomes into a highly condensed graph structure: pancontigs are connected by edges if adjacent in any input sequence, and individual genomes are represented by paths through the graph.

Here, we are applying pangraph to much shorter regions than whole genomes to arrive at a simplified representation of structural diversity in flanking regions, which are typically highly diverse and have many breaks in homology due to large-scale insertions and rearrangements. We refer to pancontigs as 'homologous blocks' of sequence and colour blocks according to this homology, using this to produce visualizations that show their coarse-grained structural arrangement. Because pangraph was developed with the Mb scale of whole genomes in mind, it can find homologous blocks within kbs of diverse flanking regions across thousands of sequences on a personal computer in minutes: for a dataset of n=1581 *bla*$_{\mathrm{TEM}}$-positive assemblies from 24 bacterial genera, building the pangraph of ~10 kb flanking regions (+/−5 kb) takes less than 3 min on a laptop (2 GHz Quad-Core Intel Core i5 processors, 16 GB RAM).

Since the aim is to provide a coarse-grained representation of homology for visualization and quantification, by default the pipeline uses a minimum block size of 100 bp with the asm10 sensitivity level in minimap2. In the pangraph representation a block has a consensus sequence and thus a consensus length, although individual sequences may have insertions or deletions, making them vary in length. This consensus sequence is used to compare and align blocks in the pangraph algorithm, which makes pangraph fast but may introduce minor inconsistencies and artefacts in the block alignments. Although we use the true lengths of sequences in all visualizations (meaning that identically coloured blocks may appear as slightly different lengths in visualizations), we do not analyse small-scale structural differences below the length scale of blocks. We note that it is possible to use this information for investigating structural rearrangement, such as using target site duplications (<15 bp direct repeats) to indicate newly transposed copies of transposable elements as in TETyper [7], but we do not investigate them here. However, no sequence information is lost and a true multiple sequence alignment for every block can be reconstructed with pangraph using the 'polish' command for more detailed analysis (see pangraph documentation).

Visualization

The pangraph of the flanking regions in GFA format can be viewed in Bandage, although for most mobile genes the default force-directed layout will produce an uninformative 'hairball' due to repeated homology blocks. We therefore produce a linear interactive html representation of the flanking region with blocks coloured by homology (Fig. 1a). Unique stretches of sequence with no homology within the dataset are coloured grey. Users can highlight a particular block by clicking on it. Optionally, if annotations are provided for the input sequences in gff format these can be projected onto this visualization and toggled on and off (for examples, see the beta-lactamase link).

Quantification

The coarse-grained representation of the flanking regions in terms of homologous blocks is our starting point for quantitative analysis. We define two quantities, illustrated schematically in Fig. 1b:

- Uninterrupted shared distance: for a pairwise comparison of two flanking regions, the distance upstream/downstream from the focal gene to their first structural difference, i.e. the first point where their paths in the graph diverge. The inverse cumulative distribution function (cdf) of all pairwise uninterrupted shared distances in a set of sequences represents the decay of structural similarity with distance from the focal gene, and can help to identify common positions where non-homologous structural variation is introduced.
- Block diversity: at a given distance away from the focal gene, the Shannon entropy of the homologous blocks at that location across the dataset. If $p_i\left(x\right)$ is the proportion of sequences at distance x with block i, then $D\left(x\right) = -\sum_i p_i\left(x\right)\log\left[p_i\left(x\right)\right]$.

Beta-lactamase dataset

We first used prevalence information from the Comprehensive Antibiotic Resistance Database (CARD) v3.1.0 [3] to assemble a dataset of genomic sequences that contained at least 1 gene encoding a beta-lactamase from any of 12 families (CMY, CTX-M, GES, IMP, KPC, NDM, OXA, PER, SHV, TEM, VIM, VEB) according to CARD's 'strict' matching criteria and were coded by CARD as 'ncbi_chromosome' (*n*=3199) or 'ncbi_plasmid' (*n*=4026). We matched information from these NCBI accessions to obtain their BioSample ID and also used NCBI entrez to link them to metadata such as species, collection date, host and geographical location (11 had no associated BioSample). We automatically assigned country names from the NCBI variable geo_loc_name and also from lat_lon where possible. We inspected 42 entries where automatic country names failed and inputted the country manually. Where samples were described as e.g. 'USA ex Mexico' we coded this as USA.

In the final cleaned metadata 5958 sequences (82.5%) had a collection year, 6215 (86.0%) had a country and 5764 (79.8%) had both. At the level of taxonomic order, most sequences were from species within *Enterobacterales* (5529/7225, 76.6%). Beta-lactamase families can be diverse and contain non-homologous genes – notably, the OXA family. Therefore, we used a review of the literature and the classification of clinically important beta-lactamase families [10] to choose 12 clinically important beta-lactamase genes that have emerged recently as mobile AMR threats as focal genes for our example quantitative analysis (Table 1). For these 12 focal genes, we only included sequences in our dataset that had sufficient flanking sequence either side of the focal gene (+/− 5 kb) with <25 nucleotide-level differences in the focal gene itself (*n*=3362 total; sequences with length <10 kb+gene length are omitted by this threshold, e.g. small plasmids). This corresponds to a nucleotide identity cutoff ranging from 96.6% for *bla*$_{\text{IMP-4}}$ (shortest gene, 741 bp) to 97.8% for *bla*$_{\text{CMY-2}}$ (longest gene, 1146 bp), although in practice nearly all included had <7 SNVs in the focal gene so were >99% identical at the nucleotide level.

We compiled this dataset in this way before we were aware of NCBI MicroBIGG-E. MicroBIGG-E allows users to query genomes already analysed with NCBI AMRFinderPlus as part of the Pathogen Detection Pipeline for a specific gene or genetic element, and then download only its flanking regions across all genomes. This makes it an ideal starting point for flanking region analysis with our pipeline. However, as of 16 November 2023 it only lets the user download flanking regions +/−2 kb. Looking at larger flanking regions currently requires downloading the whole contigs first.

APPLICATION: BETA-LACTAMASE GENES

To demonstrate the scalability of our method, we applied it to 12 different beta-lactamase genes (Table 1). Beta-lactam antibiotics are key to modern medicine, accounting for 65% of prescriptions for injectable antibiotics in the USA [11]. These antibiotics share a common component: the beta-lactam ring, first seen in the structure of penicillin. Beta-lactamases are a diverse group of enzymes that can break apart the beta-lactam ring by hydrolysis and render beta-lactam antibiotics ineffective. The use of beta-lactam antibiotics therefore exerts a strong selective pressure for sensitive bacteria to carry beta-lactamases. This effect was first observed immediately after the widespread introduction of penicillin: at one London hospital the proportion of penicillin-resistant *Staphylococcus aureus* carrying the beta-lactamase *bla*Z increased from 14% in 1946 to 38% the following year [12]. The increase in the prevalence of beta-lactamases and their adaptation to hydrolyse successive generations of beta-lactams is one of the clearest real-world examples of rapid evolution.

Table 1. Beta-lactamase genes used as focal genes for example analysis. We started from the CARD prevalence database v3.1.0 and reanalysed sequences to confirm gene presence. Chromosome/plasmid designation is taken from CARD ('ncbi_chromosome' or 'ncbi_plasmid'). Only those with sufficient flanking sequence (+/− 5 kb either side of the focal gene) with <25 nucleotide-level differences in the focal gene were taken forward for further analysis. Almost all variants were SNVs: of 6186 gene sequences extracted across the dataset, only 87 (1.5%) were not exactly the same length as the focal gene. Functional group information is as in Bush and Jacoby [10]. First descriptions are given for the specific named beta-lactamase rather than the enzyme family as a whole – for example, the CTX-M family was first described in 1990 [32]

Beta-lactamase	Group	First description	First genome	*n* (chrom/plas)	Genera
CMY-2	1, 1e	1990: Bauernfeind *et al.* [32]	2002	196 (47/149)	7
CTX-M-15	2be	2001: Karim *et al.* [33]	2001	746 (214/532)	13
CTX-M-65	2be	2009: Lee *et al.* [34]	2008	434 (90/344)	10
GES-1	2f	2000: Poirel *et al.* [35]	2008	27 (7/20)	8
IMP-4	3a	1994: Osano *et al.* [36]	1998	33 (0/33)	7
KPC-2	2f	2001: Yigit *et al.* [37]	2007	515 (17/498)	11
NDM-1	3a	2009: Yong *et al.* [38]	2009	338 (56/282)	14
OXA-10	2d	1988: Huovinen *et al.* [39]	2001	93 (24/69)	12
OXA-48	2df	2004: Poirel *et al.* [40]	2010	158 (66/92)	4
PER-1	2be	1993: Nordmann *et al.* [41]	2006	24 (16/8)	7
TEM-1	2b	1965: Datta and Kontomichalou [42]	1974	1581 (362/1212)	24
VIM-1	3a	1999: Lauretti *et al.* [43]	2001	80 (5/75)	7

Beta-lactamases are a key clinical problem in Gram-negative species [13] and recent estimates suggest much higher prevalences of beta-lactamases in the Global South [14]. The World Health Organization (WHO) has designated Gram-negative species priority pathogens for the development of new antibiotics [15]. In recent decades, many newly described beta-lactamases have been identified in clinical bacteria [13]. A particular concern are emerging extended-spectrum beta-lactamases (ESBLs), which confer resistance to a range of beta-lactams [16] and commonly spread on MGEs. In some cases these genes can be identified as having been mobilized by MGEs from the chromosomes of environmental bacteria – Partridge (2011) gives a list of the probable ancestral species for many beta-lactamases [17]. Many mobile beta-lactamases are speculated to have originated from a single mobilization event. Mobile beta-lactamases therefore provide repeated examples of a common pattern: mobilization of a chromosomal gene, followed by diversification of its flanking regions as it spreads through bacterial pangenomes into new genomic contexts under strong selective pressure.

We selected 12 clinically important beta-lactamases as focal genes and then downloaded chromosome and plasmid sequences to run through our pipeline looking at +/−5 kb flanking regions (see Methods). A single amino acid change in a beta-lactamase is enough to denote a new numbered variant (e.g. $bla_{\mathrm{NDM-1}}$ and $bla_{\mathrm{NDM-2}}$), although identical amino acids can still have synonymous SNVs in their nucleotide sequence. We arbitrarily chose 25 nucleotide-level differences in the focal gene as a cutoff to include highly related variants, approximately corresponding to a 97.5% nucleotide identity cutoff (in practice nearly all included had <7 SNVs so >99%, as expected with their recent expansion in human pathogens). Each beta-lactamase gene was seen across multiple genera and all but $bla_{\mathrm{IMP-4}}$ were found on both chromosomes and plasmids (Fig. 2).

The interactive visualizations produced are available at https://liampshaw.github.io/flanking-regions. For the remainder of this paper, we explore three general quantitative findings that hold – more or less – across these diverse genes, and comment on the possible processes that generate structural diversity.

The accumulation of nucleotide-level variation correlates with the breakdown of flanking homology

In a simple model of a gene that is mobilized from a chromosomal background and then spreads horizontally on a MGE, the accumulation of SNVs in the gene itself should be a 'slow' molecular clock in contrast to the 'fast' rearrangements that happen in its flanking regions. The two would be expected to be broadly correlated: observing more SNVs in the gene suggests that more time has elapsed, and so greater structural diversity should have accumulated in its flanking regions.

The CTX-M-9-like beta-lactamases provide a clear example of this pattern. Arbitrarily choosing the CTX-M-9-like gene $bla_{\mathrm{CTX-M-65}}$ as a focal gene and searching for sequences with variants, we find a strong correlation between SNVs in the gene and uninterrupted shared distances in flanking regions (Fig. 3). This does not appear to be due to repeated sequencing due to outbreaks, because the pattern persists if we include one random isolate from each year/country/genus combination (Fig. S1, available in the online

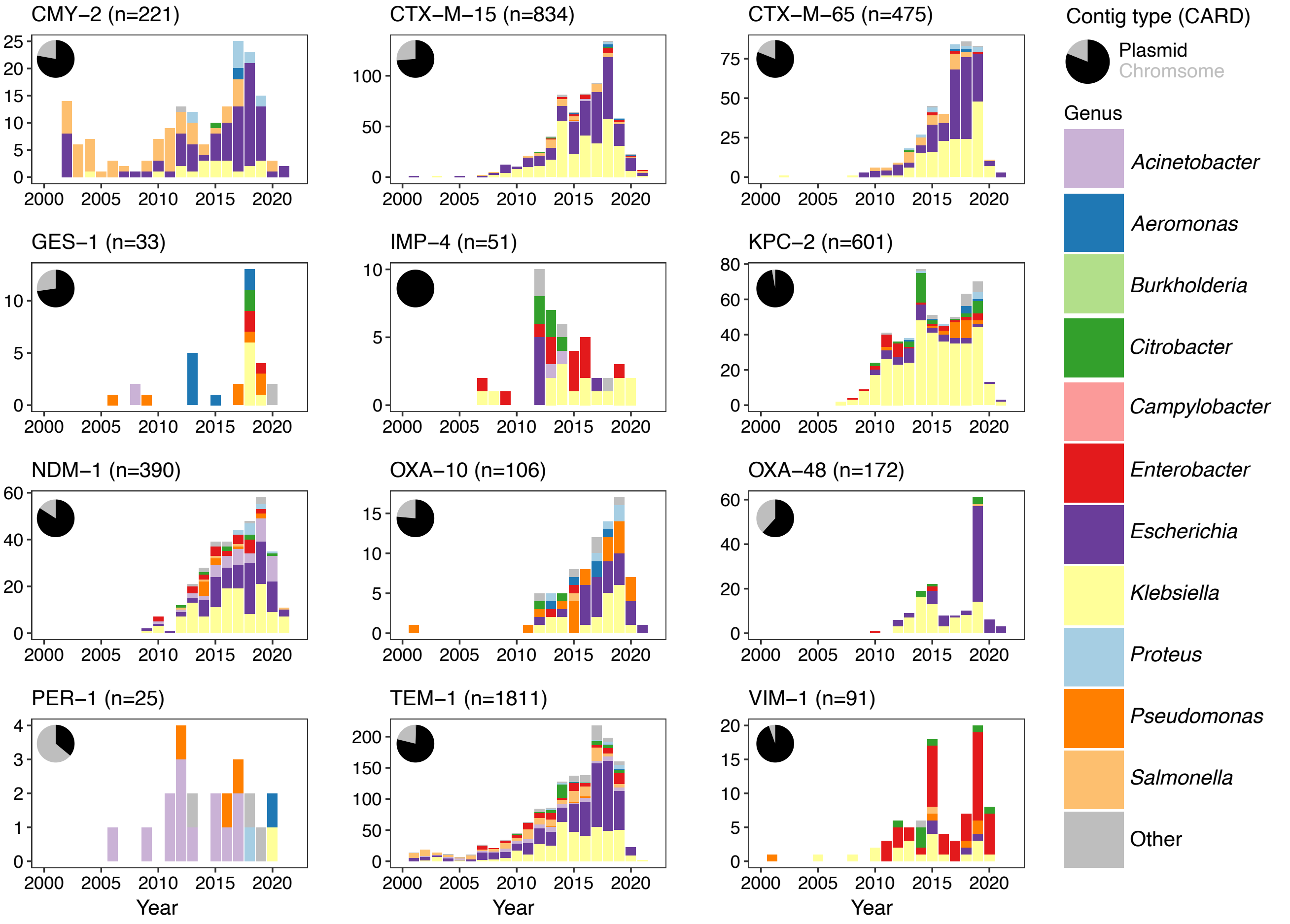

Fig. 2. Summary of the flanking regions analysed for each beta-lactamase. Inset pie charts show the original contig type as given in CARD v3.1.0 (ncbi_chromosome or ncbi_plasmid). This plot only shows sequences that passed filtering requirements (see Methods). For consistency only sequences from 2000–2022 are shown, omitting a small minority of older sequences (30 for TEM-1, 1 IMP-4). Genera counts are given for common genera, with 26 genera with <50 sequences in the original dataset grouped into 'other' (*Achromobacter*, *Alcaligenes*, *Avibacterium*, *Bacillus*, *Bacteroides*, *Chlamydia*, *Chryseobacterium*, *Cronobacter*, *Haemophilus*, *Leclercia*, *Legionella*, *Morganella*, *Mycoplasma*, *Myroides*, *Neisseria*, *Pasteurella*, *Propionibacterium*, *Providencia*, *Ralstonia*, *Raoultella*, *Serratia*, *Shewanella*, *Shigella*, *Sphingobacterium*, *Stutzerimonas*, *Vibrio*). Full genera and contig counts for each beta-lactamase are provided as Table S1.

version of this article). There is still some shared sequence around almost all CTX-M-9-like genes, supporting their common origin in some previous mobilization. Indeed, previously Olson *et al.* [18] found a chromosomal beta-lactamase in *Kluyvera georgiana* that shared 100% amino acid identity with CTX-M-14 and was in a 2.7 kb region with 99 % nucleotide identity to the complex class 1 integron In60, arguing therefore that *K. georgiana* was the likely source for the progenitor of the CTX-M-9-like group through mobilization. However, since then different CTX-M-9-like beta-lactamases have diverged in their mobilization [2].

With a model of a single initial mobilization, all subsequent disruptions of shared sequence are due to subsequent insertions (or deletions), which introduce non-homologous sequence. There are two 'modes' of this breakdown of shared sequence in the cumulative distribution of uninterrupted shared distances: gradual decay and sharp breakpoints (Fig. 3). A sharp breakpoint across pairwise comparisons might be suggestive of a consistent 'common block', which could be a coherent MGE inserting into multiple genomic backgrounds or a gene cassette in an integron's gene array; gradual decay is suggestive of a 'fossilized' MGE that is undergoing degradation via the introduction of non-homologous sequence in its flanking regions at random points. These patterns may be driven by the same underlying processes. Adjusting the pangraph parameters of the pipeline can reduce

[2] In line with common usage, by 'mobile gene' we refer to any gene that can be or has been found on MGEs in the recent past, on the scale of decades, allowing for the possibility that it may not be currently mobile in all genomes.

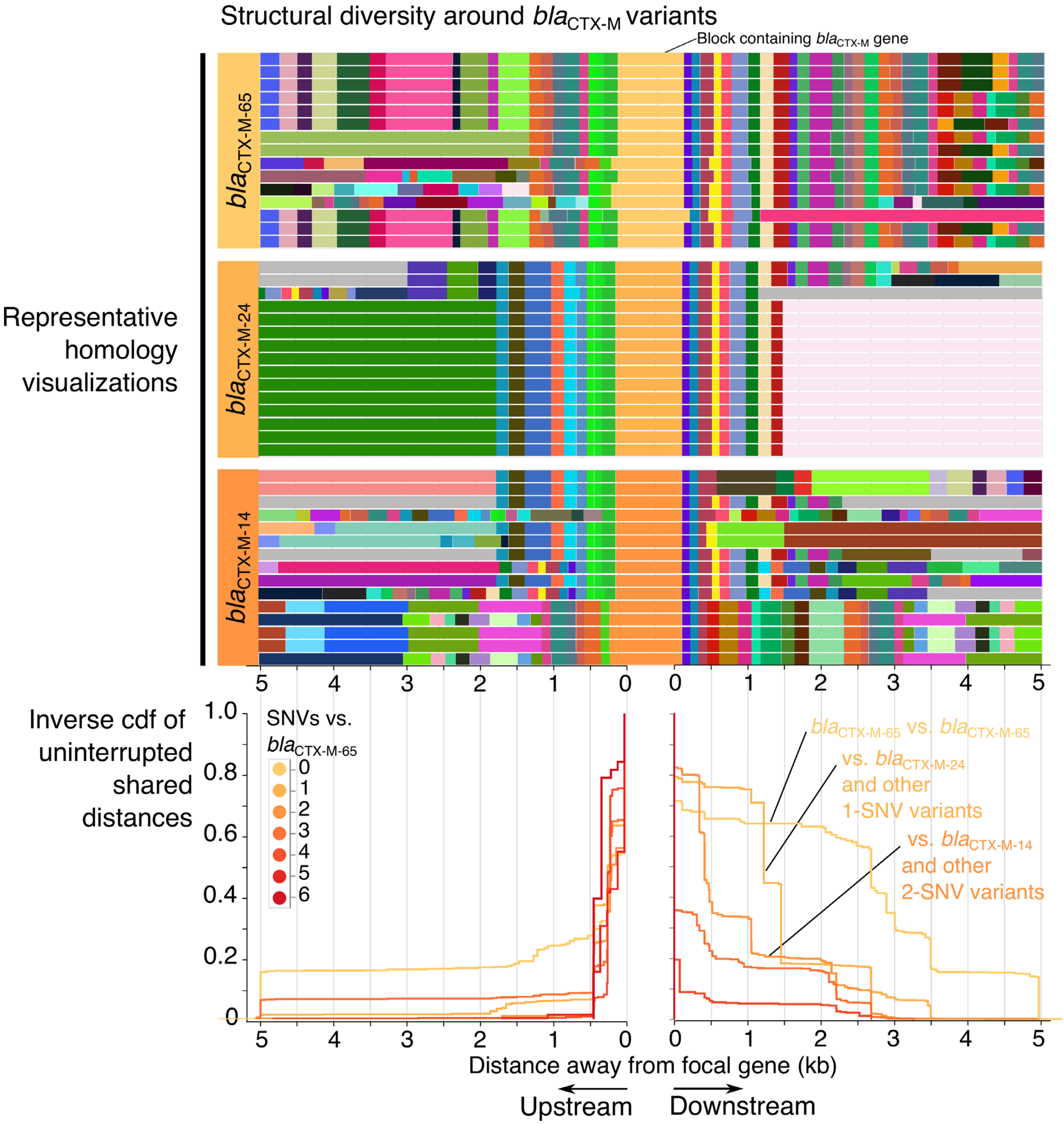

Fig. 3. The breakdown of homology in the flanking regions of *bla*$_{\text{CTX-M-65}}$and closely related genes. The upper panels show the visualization of homology in the 5 kb flanking regions of *bla*$_{\text{CTX-M-65}}$ and two closely related genes that encode different CTX-M protein variants, *bla*$_{\text{CTX-M-24}}$ (1 SNV apart) and *bla*$_{\text{CTX-M-14}}$ (2 SNVs apart). Only a subset of 15 sequences are shown for each gene. Grey blocks have no homology within the wider dataset. Below these visualizations is the inverse cumulative distribution function of pairwise comparisons of distance to the first breakpoint (first occurrence of non-homologous sequence) between n=434 sequences carrying any *bla*$_{\text{CTX-M}}$ gene with <25 SNVs to *bla*$_{\text{CTX-M-65}}$, stratified by the number of SNVs in the focal gene relative to *bla*$_{\text{CTX-M-65}}$ (only comparisons involving *bla*$_{\text{CTX-M-65}}$ are shown). The same pattern is seen when picking a single isolate for each year/country/genus combination to control for potential sampling bias (Fig. S1).

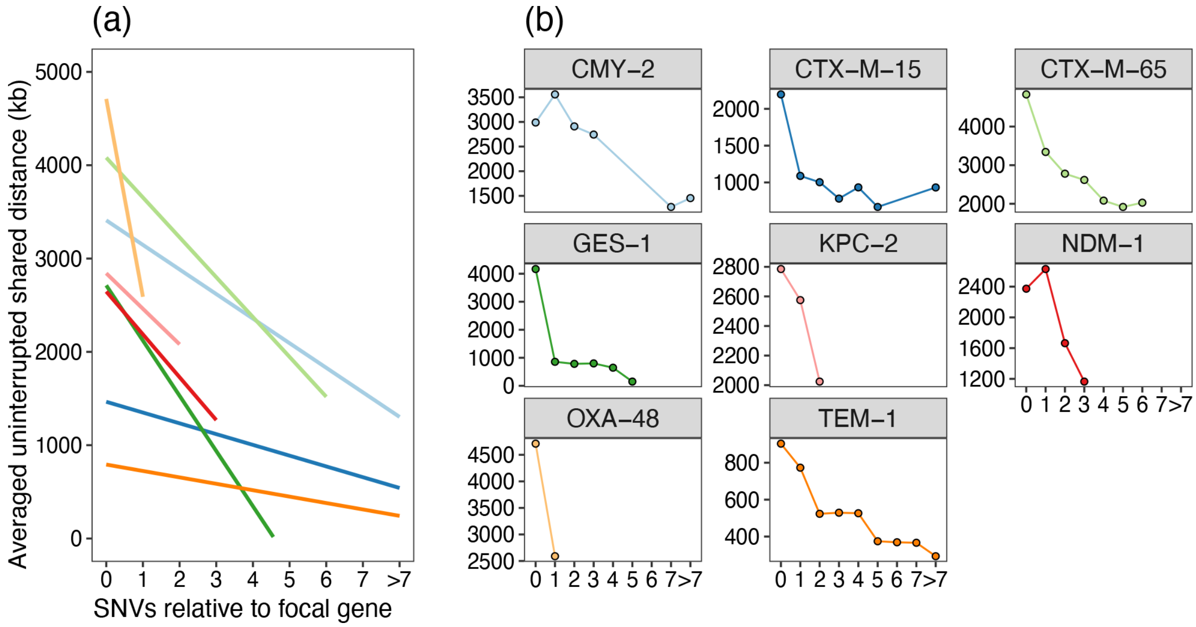

Fig. 4. Structural diversity correlates with nucleotide level variation. Across different beta-lactamases the presence of any SNV in the focal gene is generally correlated with a shorter flanking region overlap (sum of average upstream and downstream pairwise uninterrupted shared distances). (a) Linear fits (stat_smooth) and (b) average per SNV level comparison for *n*=7 genes with >100 sequences in the initial dataset (Table 1). Not all genes have variants in the dataset for all possible values of SNVs, so where points are absent in (b) this denotes an absence of variants at this difference level. Pairwise comparisons have been deduplicated for year/country/genus combinations before plotting. The flanking region overlap is the sum of the upstream and downstream average uninterrupted shared distances. Number of data points for each gene (after deduplicating): *bla*$_{\text{CMY−2}}$ (*n*=108), *bla*$_{\text{CTX−M−15}}$ (*n*=274), *bla*$_{\text{CTX}}$−M−65 (*n*=119), *bla*$_{\text{GES}}$−1 (*n*=15), *bla*$_{\text{KPC}}$−2 (*n*=161), *bla*$_{\text{NDM}}$−1 (*n*=175), *bla*$_{\text{OXA}}$−48 (*n*=66) and *bla*$_{\text{TEM}}$−1 (*n*=514).

or increase the sensitivity to non-homologous sequence. Our experience suggests that using the pipeline with default parameters does well at capturing breaks that, when inspected, are clearly caused by ‘large’ insertions/deletions of sequence (i.e. >100 bp).

The area under the curve (AUC) of the distribution of uninterrupted shared distances for shared sequence between pairs of sequences is equal to the average uninterrupted shared distance length. We can therefore use this to quantify the rough correlation between nucleotide level variation and structural diversity. By stratifying pairwise comparisons between sequences based on the SNVs in the focal gene, there is a tendency for greater diversification around the focal gene with more SNVs (Fig. 4).

Linking structural diversity to annotations

It is well established that insertion sequences (ISs), which use transposases to catalyse DNA cleavage and strand transfer leading to movement within genomes, are fundamental to the mobilization of AMR genes and contribute to the complexity of AMR flanking regions. Repeated ISs are often the reason why flanking regions cannot be assembled in genome assemblies. We used existing annotations from the NCBI gff to identify the locations of ISs as a proxy for IS presence, calculating the transposase density at a position as the number of annotations at that position that contained the term ‘transposase’ divided by the total number of sequences.

Plotting the block diversity and the uninterrupted shared distance distribution with the transposase density revealed interesting patterns (Fig. S2). As an example: for *bla*$_{\text{NDM−1}}$, the highest density of transposases is upstream, with an associated immediate breakdown of shared sequence (Fig. 5). After this breakdown of homology, the block diversity is high, indicating that the gene has become stabilized in many different backgrounds. Downstream, the breakdown of the shared ancestral background is more gradual. This recapitulates the more detailed analysis of Acman *et al.* [4].

A bias for greater structural diversity upstream

Using the distribution of uninterrupted shared distances, we found evidence for an asymmetry between the upstream and downstream flanking regions. After removing genes that are known to be found on gene cassettes and associated with integrons (*bla*$_{\text{GES-1}}$, *bla*$_{\text{IMP-4}}$, *bla*$_{\text{OXA-10}}$ and *bla*$_{\text{VIM-1}}$; see [19]), most remaining genes lie below the line of equality, meaning a greater conservation of the downstream flanking region compared to greater structural diversity upstream (Fig. 6). This suggests that for mobile beta-lactamases associated with ISs, there is a bias for a higher realized rate of introduction of new sequence into the upstream

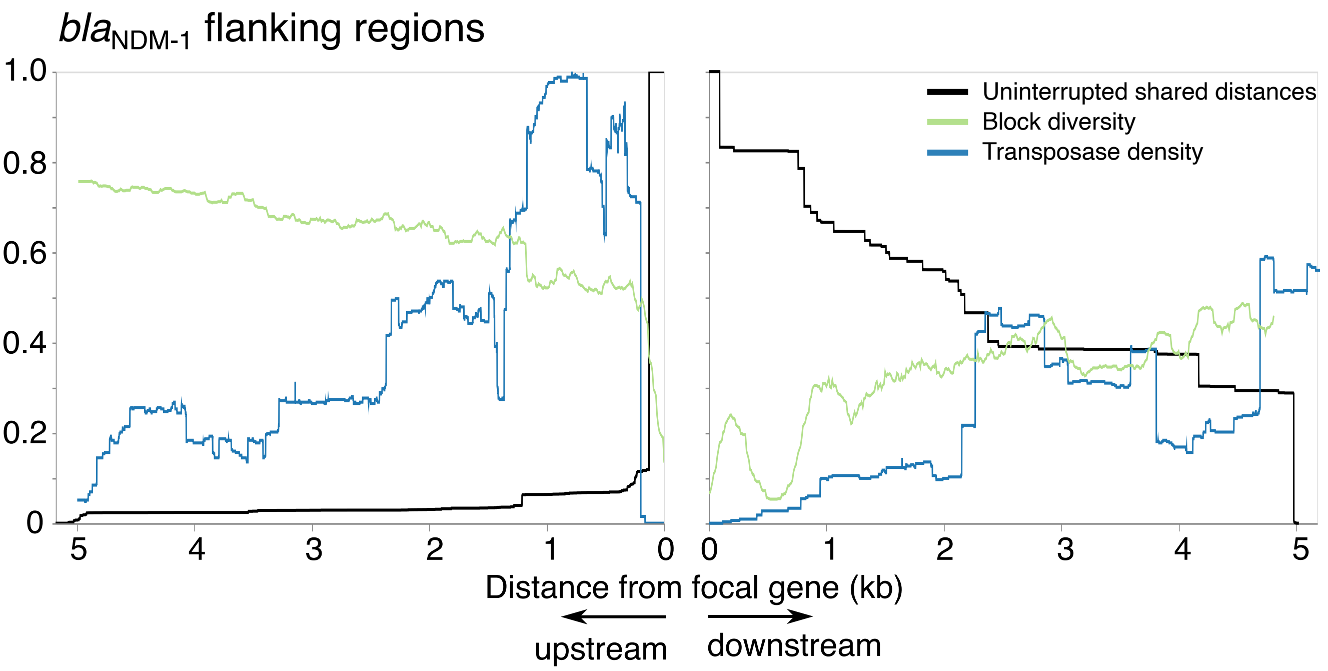

Fig. 5. Homology decay, block diversity and transposase density around *bla*$_{\text{NDM-1}}$. Homology decay in terms of pairwise distance to first breakpoint (black), block diversity (green) and transposase density from annotations as a proxy for IS density (blue). All curves are normalized (block diversity to log(*n*), transposase density to the point of maximum density in the flanking region shown (290/333 sequences). Left shows upstream and right shows downstream.

region. Although the numbers are small, it is interesting to note that the exceptions are the oldest genes in the dataset: *bla*$_{\text{TEM-1}}$ (first described in 1964) and *bla*$_{\text{CMY-2}}$ (1990). The next most recent gene, *bla*$_{\text{PER-1}}$ (1993) lies close to the line of equality, with the five remaining genes all described in the 2000s. From this limited evidence, we speculate that this difference between upstream and downstream could be one that is strongest during initial IS-associated spread, fading as genes are recruited into multi-resistance regions.

DISCUSSION

We have aimed to provide a starting point to quantitatively investigate structural diversity around mobile genes. By using pangraph to find homologous sequences in flanking regions without requiring annotation information, we have constructed a pipeline that quickly produces interactive visualizations that can be explored by researchers to make sense of these complex regions, as well as computing some basic summary statistics. To demonstrate our pipeline, we applied it to the flanking regions of 12 different betalactamase genes in public assemblies. Our observations recapitulate previous knowledge about individual beta-lactamase genes, but at scale, and provide evidence of general patterns.

We wish to highlight three limitations of our approach. First, we stress that while this approach provides a way to quantify structural diversity around a gene, it is not intended to infer how a specific instance of the gene is *currently* mobilized. For simplicity we have analysed 5 kb flanking regions, but these are unlikely to capture all information on mobilization: in Gram-negative species mobilized genes tend to cluster in 'multiresistance regions', which can be tens of kilobases in size [20]. Second, our approach is coarse-grained – in the sense that pangraph has a minimum size limit for the size of homologous blocks. This means that traces of evolution at smaller scales will be missed by considering homologous blocks as 'identical', for example transposition events which are associated with small-scale (2 bp) insertions. More fine-grained analysis is possible using the multiple-sequence alignments that are generated for each block by pangraph, although a bespoke analysis (e.g. an alignment against a known structure to identify small-scale changes) would always be expected to be superior. In particular, an issue in pangraph is that breaks in homology involving small blocks close to the minimum size (100 bp) appear to be more unreliable, and we are investigating how best to process these. Third, our analysis uses public genomes that are a biased sample, so all inferences about evolutionary events should be viewed with appropriate scepticism.

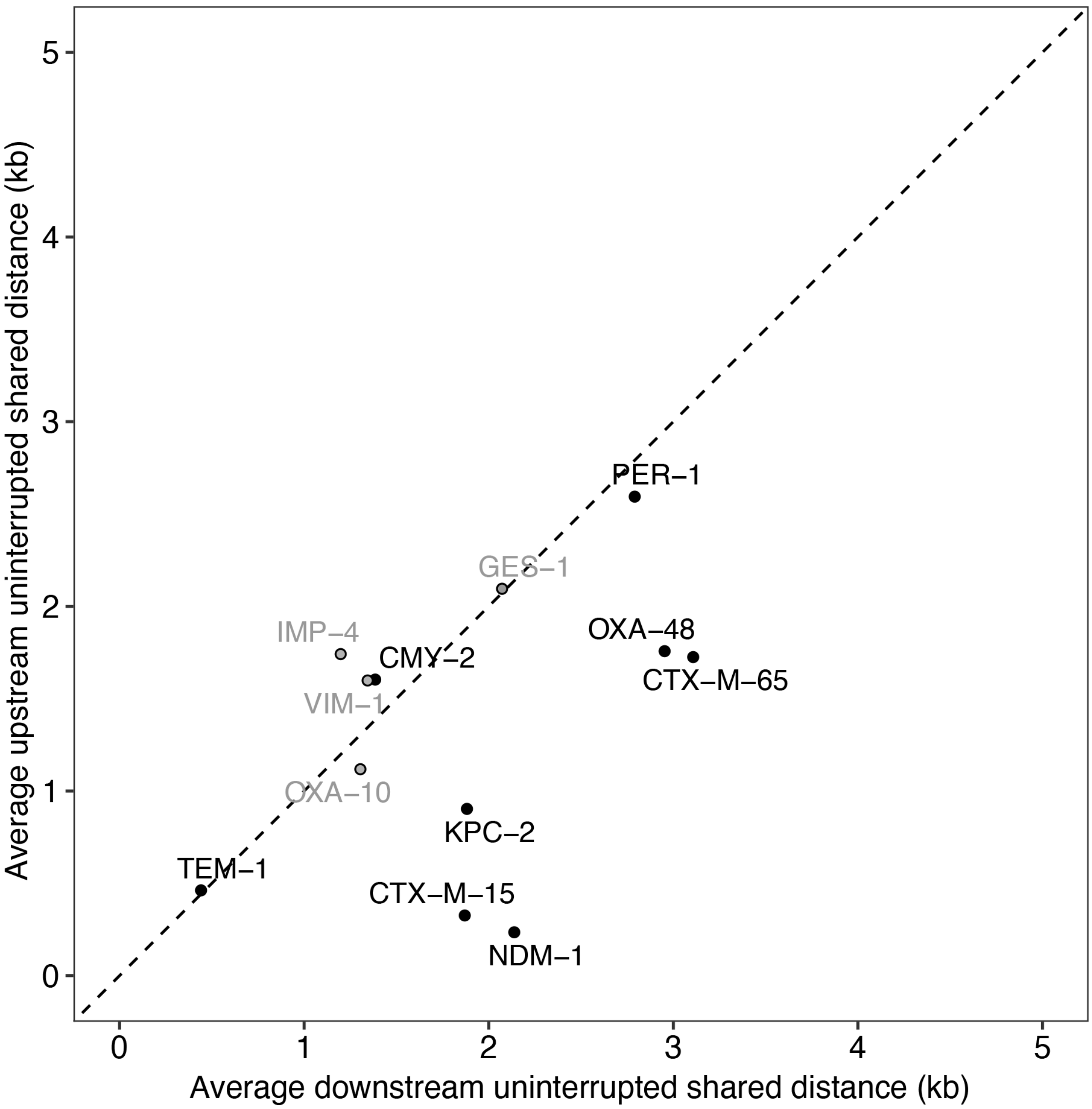

Fig. 6. Comparing upstream and downstream uninterrupted shared distances. Results show average downstream and upstream pairwise uninterrupted shared distances for each focal gene, only including comparisons between identical focal genes (0 SNVs) after metadata deduplication for year/genus/country. Although most genes lie on the line of equality, there seems to be a bias in some for greater conservation downstream. There is no overall significant difference at α=0.05 (7/12 downstream>upstream; Wilcoxon signed-rank exact test V=57, *P*=0.18). However, when integron-associated genes (grey) are removed, the difference becomes stronger (6/8 downstream>upstream, V=32, *P*=0.055).

We are far from being able to build a quantitative evolutionary model of the structural diversity around mobile genes. However, the patterns we find for beta-lactamases suggest some general principles about the evolution of flanking regions, which appear to be consistent with what has previously been noted in the literature.

First, nucleotide-level variation in the mobile gene itself does indeed appear to be a ‘slow’ molecular clock compared to the much faster rates of rearrangements in flanking regions. Even single SNVs can change the hydrolytic profile of beta-lactamases, for example in the well-studied evolution of $bla_{\text{TEM-1}}$ [21, 22], including specifically while on MGEs [23, 24], so this variation should not generally be taken as a neutral clock. However, this nevertheless suggests a quantitative connection between mutational and non-mutational evolution that could be fruitful for further exploration. Second, we confirm that ISs can be linked to high levels of structural diversity around mobile genes, likely driven by homologous recombination as well as transposition. Third, we found that some genes had an asymmetry with direction, with greater conservation of downstream flanking regions in the majority of non-integron-associated beta-lactamases (6/8), meaning greater structural diversity upstream.

Speculatively, we propose that this last observation could be a signature of the mobilization dynamics of IS-associated beta-lactamases. It is known that many beta-lactamases have been mobilized from chromosomal backgrounds and are weakly expressed in their native genomic context. A wide body of evidence suggests that there is a deep connection between mobility and expression. Not only has antibiotic treatment been shown experimentally to select for the movement of transposon-associated genes from chromosomes to plasmids [25], but it was noted as far back as the 1990s that the inverted repeats around ISs often contain promoters, and that such ISs are often found immediately upstream of beta-lactamase genes, e.g. for $bla_{\mathrm{TEM-6}}$ [26] or $bla_{\mathrm{CTX-M-14}}$ [27]. For instance, Poirel *et al*. observed that the upstream insertion sequence ISEcp1, which mobilizes $bla_{\mathrm{CTX-M-19}}$ and other resistance genes, contains a strong promoter, simultaneously mobilizing the resistance gene and providing strong expression [28]. The spread of beta-lactamases is due to strong selection for increased expression: ISs contribute to this in two ways, by providing both a means of transfer onto plasmids and a strong promoter.

Thinking with this simple conceptual model, the initial mobilization of an ancestral beta-lactamase from a chromosome is likely to require the upstream insertion of at least one IS. Therefore, that upstream location will then be a hotspot for the creation of new structural diversity, whether through subsequent insertion to the same target site or homologous recombination between common regions of ISs. We observed that where beta-lactamases are found in gene cassettes and known to be integron-associated (which is the case for $bla_{\mathrm{IMP-4}}$, $bla_{\mathrm{OXA-10}}$, $bla_{\mathrm{VIM-1}}$, $bla_{\mathrm{GES-1}}$), we did not observe this asymmetry. We also observed that more recent beta-lactamases (described in the 2000s) had more diversity upstream than older genes, suggesting that this trend may decay over time as the gene becomes more embroiled in other genetic contexts, with the initial era of spread dominated by a single ancestral MGE coming to an end. Indeed, over time AMR genes tend to accumulate into multi-resistance regions, which has been suggested to be driven by homologous recombination between common components such as transposases [29]. Repeated rounds of homologous recombination and insertion that degrade 'active' mobile MGEs may play a similar role in the creation of multi-resistance regions to that suggested to be responsible for the creation of defence islands in chromosomes [30].

SUMMARY

The approach we outline for visualizing and quantifying structural diversity in flanking regions is applicable to any gene and scales to thousands of sequences on the kb scale. We have applied it to recently emerged beta-lactamases as an example of mobile genes. Each mobile gene is worth detailed study on its own and homology visualizations can help understand the patterns in its flanking regions. Large-scale analysis across genes can reveal patterns consistent with similar underlying processes, suggesting conclusions consistent with the existing literature. Future quantitative methods will allow us to better understand the dynamics that govern the arrival and establishment of genes within pangenomes.

Funding information
L.P.S. is a Sir Henry Wellcome Postdoctoral Fellow funded by Wellcome (grant 220422/Z/20/Z). R.A.N. is funded by the University of Basel.

Acknowledgements
The authors thank Sally Partridge, Zamin Iqbal and Will Matlock for helpful feedback on a draft version, and Marco Molari for ongoing and stimulating discussions. L.P.S. thanks John Lees and Craig MacLean for feedback on visualizations.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, *et al*. Characterization of a new metallo-beta-lactamase gene, bla(NDM-1), and a novel erythromycin esterase gene carried on a unique genetic structure in *Klebsiella pneumoniae* sequence type 14 from India. *Antimicrob Agents Chemother* 2009;53:5046–5054.
2. Jones LS, Toleman MA, Weeks JL, Howe RA, Walsh TR, *et al*. Plasmid carriage of bla NDM-1 in clinical *Acinetobacter baumannii* isolates from India. *Antimicrob Agents Chemother* 2014;58:4211–4213.
3. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, *et al*. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res* 2023;51:D690–D699.
4. Acman M, Wang R, van Dorp L, Shaw LP, Wang Q, *et al*. Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene bla_{NDM} *Nat Commun* 2022;13:1131.
5. Toleman MA, Spencer J, Jones L, Walsh TR. blaNDM-1 is a chimera likely constructed in *Acinetobacter baumannii*. *Antimicrob Agents Chemother* 2012;56:2773–2776.
6. Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* 2021;37:2473–2475.
7. 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* 2018;4:12.
8. Matlock W, Lipworth S, Constantinides B, Peto TEA, Walker AS, *et al*. Flanker: a tool for comparative genomics of gene flanking regions. *Microb Genom* 2021;7:000634.
9. Noll N, Molari M, Shaw LP, Neher RA. PanGraph: scalable bacterial pan-genome graph construction. *Microb Genom* 2023;9.
10. Bush K, Jacoby GA. Updated functional classification of β-lactamases. *Antimicrob Agents Chemother* 2010;54:969–976.
11. Bush K, Bradford PA. β-lactams and β-lactamase inhibitors: an overview. *Cold Spring Harb Perspect Med* 2016;6:a025247.
12. Barber M. Staphylococcal infection due to penicillin-resistant strains. *Br Med J* 1947;2:863–865.

13. Bush K, Bradford PA. Epidemiology of β-lactamase-producing pathogens. *Clin Microbiol Rev* 2020;33:e00047-19.
14. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 2022;399:629–655.
15. World Health Organization. Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis; 2017. https://www.who.int/publications/i/item/WHO-EMP-IAU-2017.12
16. Livermore DM. Defining an extended-spectrum beta-lactamase. *Clin Microbiol Infect* 2008;14 Suppl 1:3–10.
17. Partridge SR. Analysis of antibiotic resistance regions in Gram-negative bacteria. *FEMS Microbiol Rev* 2011;35:820–855.
18. Olson AB, Silverman M, Boyd DA, McGeer A, Willey BM, *et al*. Identification of a progenitor of the CTX-M-9 group of extended-spectrum beta-lactamases from Kluyvera georgiana isolated in Guyana. *Antimicrob Agents Chemother* 2005;49:2112–2115.
19. Partridge SR, Tsafnat G, Coiera E, Iredell JR. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev* 2009;33:757–784.
20. Partridge SR, Enne VI, Grohmann E, Hall RM, Rood JI, *et al*. Classifying mobile genetic elements and their interactions from sequence data: the importance of existing biological knowledge. *Proc Natl Acad Sci U S A* 2021;118:35.
21. Barlow M, Hall BG. Predicting evolutionary potential: in vitro evolution accurately reproduces natural evolution of the tem beta-lactamase. *Genetics* 2002;160:823–832.
22. Weinreich DM, Delaney NF, Depristo MA, Hartl DL. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 2006;312:111–114.
23. Kosterlitz O, Grassi N, Werner B, McGee RS, Top EM, *et al*. Evolutionary “Crowdsourcing”: alignment of fitness landscapes allows for cross-species adaptation of a horizontally transferred gene. *Mol Biol Evol* 2023;40:msad237.
24. Rodriguez-Beltran J, Hernandez-Beltran JCR, DelaFuente J, Escudero JA, Fuentes-Hernandez A, *et al*. Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation. *Nat Ecol Evol* 2018;2:873–881.
25. Yao Y, Maddamsetti R, Weiss A, Ha Y, Wang T, *et al*. Intra- and inter-population transposition of mobile genetic elements driven by antibiotic selection. *Nat Ecol Evol* 2022;6:555–564.
26. Goussard S, Sougakoff W, Mabilat C, Bauernfeind A, Courvalin P. An IS1-like element is responsible for high-level synthesis of extended-spectrum beta-lactamase TEM-6 in Enterobacteriaceae. *J Gen Microbiol* 1991;137:2681–2687.
27. Cao V, Lambert T, Courvalin P. ColE1-like plasmid pIP843 of *Klebsiella pneumoniae* encoding extended-spectrum beta-lactamase CTX-M-17. *Antimicrob Agents Chemother* 2002;46:1212–1217.
28. Poirel L, Decousser J-W, Nordmann P. Insertion sequence ISEcp1B is involved in expression and mobilization of a bla(CTX-M) beta-lactamase gene. *Antimicrob Agents Chemother* 2003;47:2938–2945.
29. Partridge SR, Zong Z, Iredell JR. Recombination in IS26 and Tn2 in the evolution of multiresistance regions carrying blaCTX-M-15 on conjugative IncF plasmids from *Escherichia coli*. *Antimicrob Agents Chemother* 2011;55:4971–4978.
30. Rocha EPC, Bikard D. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLOS Biol* 2022;20:e3001514.
31. Navarro F, Mesa R-J, Miró E, Gómez L, Mirelis B, *et al*. Evidence for convergent evolution of CTX-M-14 ESBL in *Escherichia coli* and its prevalence. *FEMS Microbiol Lett* 2007;273:120–123.
32. Bauernfeind A, Jungwirth R, Schweighart S, Theopold M. Antibacterial activity and beta-lactamase stability of eleven oral cephalosporins. *Infection* 1990;18 Suppl 3:S155–67.
33. Karim A, Poirel L, Nagarajan S, Nordmann P. Plasmid-mediated extended-spectrum beta-lactamase (CTX-M-3 like) from India and gene association with insertion sequence ISEcp1. *FEMS Microbiol Lett* 2001;201:237–241.
34. Lee SG, Jeong SH, Lee H, Kim CK, Lee Y, *et al*. Spread of CTX-M-type extended-spectrum beta-lactamases among bloodstream isolates of *Escherichia coli* and *Klebsiella pneumoniae* from a Korean hospital. *Diagn Microbiol Infect Dis* 2009;63:76–80.
35. Poirel L, Le Thomas I, Naas T, Karim A, Nordmann P. Biochemical sequence analyses of GES-1, A novel class A extended-spectrum beta-lactamase, and the class 1 integron In52 from *Klebsiella pneumoniae*. *Antimicrob Agents Chemother* 2000;44:622–632.
36. Osano E, Arakawa Y, Wacharotayankun R, Ohta M, Horii T, *et al*. Molecular characterization of an enterobacterial metallo beta-lactamase found in a clinical isolate of Serratia marcescens that shows imipenem resistance. *Antimicrob Agents Chemother* 1994;38:71–78.
37. Yigit H, Queenan AM, Anderson GJ, Domenech-Sanchez A, Biddle JW, *et al*. Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of *Klebsiella pneumoniae*. *Antimicrob Agents Chemother* 2001;45:1151–1161.
38. Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, *et al*. Characterization of a new metallo-beta-lactamase gene, bla(NDM-1), and a novel erythromycin esterase gene carried on a unique genetic structure in *Klebsiella pneumoniae* sequence type 14 from India. *Antimicrob Agents Chemother* 2009;53:5046–5054.
39. Huovinen P, Huovinen S, Jacoby GA. Sequence of PSE-2 beta-lactamase. *Antimicrob Agents Chemother* 1988;32:134–136.
40. Poirel L, Héritier C, Tolün V, Nordmann P. Emergence of oxacillinase-mediated resistance to imipenem in *Klebsiella pneumoniae* . *Antimicrob Agents Chemother* 2004;48:15–22.
41. Nordmann P, Ronco E, Naas T, Duport C, Michel-Briand Y, *et al*. Characterization of a novel extended-spectrum beta-lactamase from *Pseudomonas aeruginosa*. *Antimicrob Agents Chemother* 1993;37:962–969.
42. Datta N, Kontomichalou P. Penicillinase synthesis controlled by infectious R factors in Enterobacteriaceae. *Nature* 1965;208:239–241.
43. Lauretti L, Riccio ML, Mazzariol A, Cornaglia G, Amicosante G, *et al*. Cloning and characterization of blaVIM, a new integron-borne metallo-beta-Lactamase gene from a *Pseudomonas aeruginosa* clinical isolate. *Antimicrob Agents Chemother (Bethesda)* 1999;43:1584–1590.