

# TETyper: a bioinformatic pipeline for classifying variation and genetic contexts of transposable elements from short-read whole-genome sequencing data

Anna E. Sheppard,<sup>1,2,\*</sup> Nicole Stoesser,<sup>1</sup> Ian German-Mesner,<sup>3</sup> Kasi Vegesana,<sup>3</sup> A. Sarah Walker,<sup>1</sup> Derrick W. Crook<sup>1</sup> and Amy J. Mathers<sup>4,5</sup>

## Abstract

Much of the worldwide dissemination of antibiotic resistance has been driven by resistance gene associations with mobile genetic elements (MGEs), such as plasmids and transposons. Although increasing, our understanding of resistance spread remains relatively limited, as methods for tracking mobile resistance genes through multiple species, strains and plasmids are lacking. We have developed a bioinformatic pipeline for tracking variation within, and mobility of, specific transposable elements (TEs), such as transposons carrying antibiotic-resistance genes. TETyper takes short-read whole-genome sequencing data as input and identifies single-nucleotide mutations and deletions within the TE of interest, to enable tracking of specific sequence variants, as well as the surrounding genetic context(s), to enable identification of transposition events. A major advantage of TETyper over previous methods is that it does not require a genome reference. To investigate global dissemination of *Klebsiella pneumoniae* carbapenemase (KPC) and its associated transposon Tn4401, we applied TETyper to a collection of over 3000 publicly available Illumina datasets containing *bla*<sub>KPC</sub>. This revealed surprising diversity, with over 200 distinct flanking genetic contexts for Tn4401, indicating high levels of transposition. Integration of sample metadata revealed insights into associations between geographic locations, host species, Tn4401 sequence variants and flanking genetic contexts. To demonstrate the ability of TETyper to cope with high-copy-number TEs and to track specific short-term evolutionary changes, we also applied it to the insertion sequence IS26 within a defined *K. pneumoniae* outbreak. TETyper is implemented in python and is freely available at <https://github.com/aesheppard/TETyper>.

## DATA SUMMARY

1. TETyper source code has been deposited in GitHub (url – <https://github.com/aesheppard/TETyper>).

## INTRODUCTION

Increasing antibiotic resistance in a range of bacterial pathogens is a major global health threat, but our understanding of resistance gene dissemination remains incomplete. Many resistance genes are carried on mobile genetic elements (MGEs), enabling bacteria to evolve in response to antimicrobial pressures via gene exchange.

MGEs of relevance include plasmids, which are extra-chromosomal, usually circular, DNA structures that can be transferred between different host bacteria, and transposable elements (TEs), which are short stretches of DNA, often carried on plasmids, that can autonomously mobilise to new genomic locations via transposition [1–3]. TEs comprise transposons, which carry additional cargo genes, such as antibiotic-resistance genes, and insertion sequences (ISs), which comprise only elements necessary for transposition; however, composite structures involving ISs can also be involved in resistance gene mobilisation [4, 5].

Received 9 April 2018; Accepted 12 October 2018

**Author affiliations:** <sup>1</sup>Nuffield Department of Medicine, University of Oxford, Oxford, UK; <sup>2</sup>National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, UK; <sup>3</sup>Health Information & Technology, University of Virginia Health System, Charlottesville, Virginia, USA; <sup>4</sup>Division of Infectious Disease and International Health, Department of Medicine, University of Virginia Health System, Charlottesville, Virginia, USA; <sup>5</sup>Clinical Microbiology Laboratory, Department of Pathology, University of Virginia Health System, Charlottesville, Virginia, USA.

**\*Correspondence:** Anna E. Sheppard, [anna.sheppard@ndm.ox.ac.uk](mailto:anna.sheppard@ndm.ox.ac.uk)

**Keywords:** antimicrobial resistance (AMR); mobile genetic element (MGE); transposable element (TE); transposon; whole-genome sequencing (WGS); *Klebsiella pneumoniae* carbapenemase (KPC).

**Abbreviations:** IS, insertion sequence; KPC, *Klebsiella pneumoniae* carbapenemase; MGE, mobile genetic element; SNV, single nucleotide variant; TE, transposable element; TSD, target site duplication; WGS, whole-genome sequencing.

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

Whole-genome sequencing (WGS) has revolutionised the analysis of pathogen transmission by enabling high-resolution insight into chromosomal relatedness [6–9]. However, resistance gene dissemination via MGEs is more complicated because horizontal transfer disrupts pairing between resistance genes and host strains. To assess relatedness from the perspective of a mobile resistance gene, it is necessary to examine the gene's genetic context. However, the most widely used WGS technologies (e.g. Illumina) produce short sequencing reads; these result in fragmented assemblies, with resistance genes often present on very short contigs due to associations with repetitive elements such as TEs. This makes tracking the associated plasmids largely impractical, as assembling complete plasmid sequences from short reads is problematic [10], and reference-based approaches can be unreliable due to transposition or homologous recombination disrupting pairing between host plasmids and resistance genes [11]. Alternatively, if a resistance gene has a stable association with a specific transposon, then tracking the transposon may be a better proxy for understanding resistance gene dissemination.

One example of such an association is the *Klebsiella pneumoniae* carbapenemase (KPC) gene *bla*<sub>KPC</sub> and its associated replicative transposon Tn4401. *bla*<sub>KPC</sub> was first identified in the USA in 1996, and has since spread globally, being responsible for a large proportion of carbapenem-resistant Enterobacteriaceae infections worldwide [12–15]. Initially, it was largely associated with *K. pneumoniae* multi-locus sequence type (ST) 258 and the IncFII plasmid pKpQIL, but it has since spread to various other plasmids, other *K. pneumoniae* strains, other species of Enterobacteriaceae, and occasionally non-Enterobacteriaceae [11, 16–18]. Given the importance of Tn4401 transposition in facilitating this spread, the ability to track transposition events and sequence variation within Tn4401 may be helpful for better understanding *bla*<sub>KPC</sub> dissemination.

A similar tracking approach may also be useful for investigating the evolution of other TEs of interest. For example, replicative intermolecular transposition (where the target site for transposition is from a different DNA molecule) of a variety of TEs, including Tn4401, involves target site duplication (TSD). This results in short (~2–14 bp) direct repeats flanking the newly transposed copy [3]. Intramolecular transposition (where the transposition target site is part of the same DNA molecule) disrupts these flanking repeats, and investigation of TSD sequences can be used to gain insight into historical transposition events, as has been demonstrated for the widely dispersed IS26 [19].

Most available tools for characterising TE mobilisation from WGS data rely on detecting variation relative to a reference sequence (e.g. [20–23]). While this may be appropriate for eukaryotic (and some prokaryotic) genomes, bacterial TEs are often carried on plasmids for which no suitable references are available. ISMapper [20] was developed specifically for detecting TE insertion sites in bacterial genomes, and can be run in two modes, using either a reference genome as

## IMPACT STATEMENT

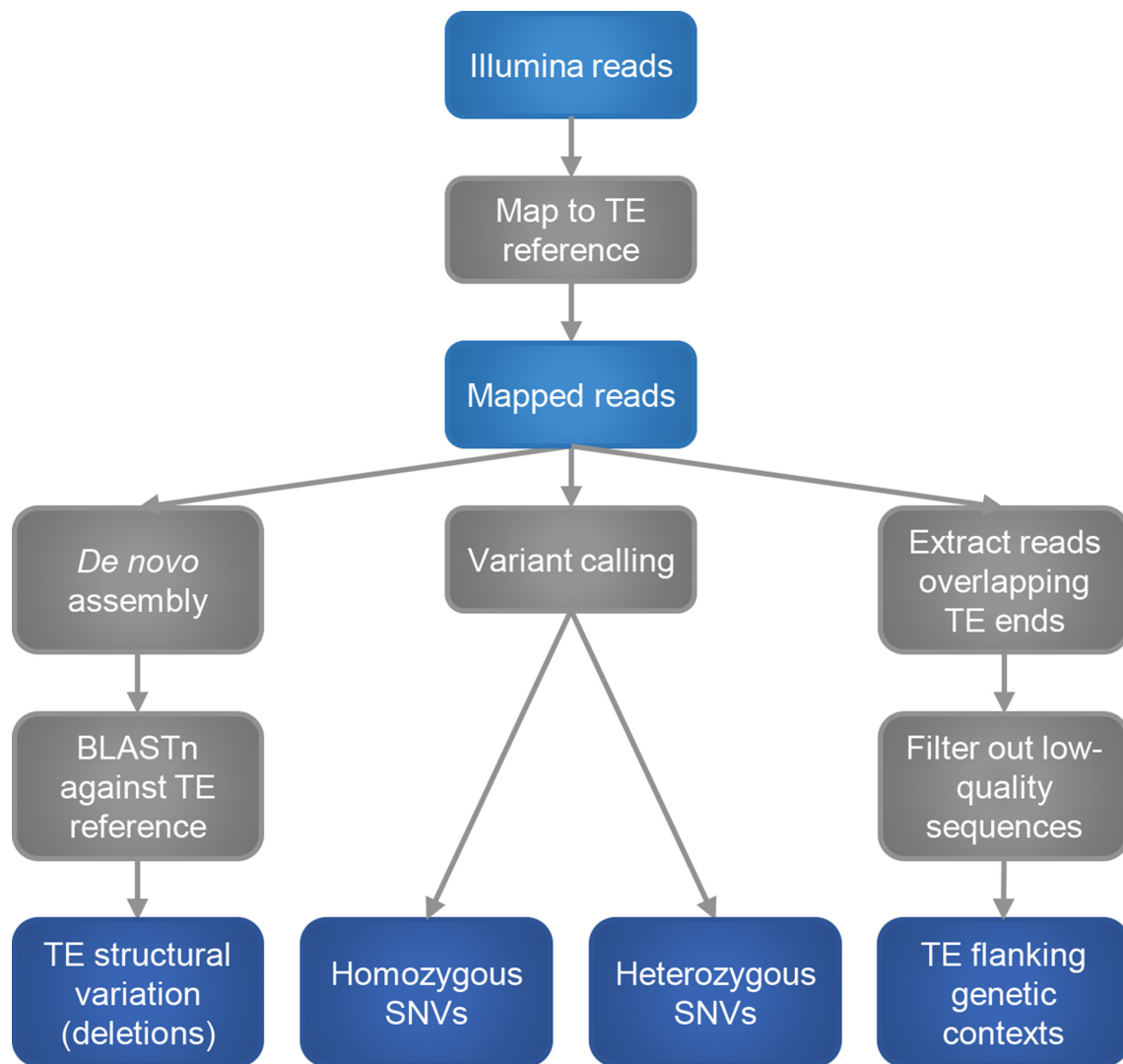
Whole-genome sequencing (WGS) of bacterial pathogens has revolutionised the analysis of global and within-outbreak transmission pathways. However, the study of antibiotic resistance dissemination is more challenging, as resistance genes are often associated with mobile genetic elements (MGEs) that enable gene exchange between different host bacteria. Therefore, standard WGS approaches that focus on host strain relationships may not be informative for understanding resistance gene dissemination. We have developed a bioinformatic tool for analysing WGS data from the perspective of a specific MGE–resistance gene association. The outputs produced identify variation within the MGE, as well as signatures of MGE mobility. This information can then be used to track the movement of the resistance gene, thus overcoming previous limitations by defining relationships from a resistance gene perspective, rather than a host-strain perspective. In an epidemiological context, this can provide insight into specific transmission pathways, thus informing infection control within outbreak scenarios, as well as increasing our understanding of global pathways of resistance dissemination.

input, or a *de novo* assembly of the sample of interest. In the latter mode, it predicts which contigs are linked to the TE of interest, but does not provide detailed information on sequence variation within the TE or base-pair resolution of insertion sites. To provide a tool for characterising TE insertion sites and sequence variation from short-read WGS data, without relying on a genome reference, and producing easily comparable outputs for epidemiological tracking purposes, we developed TETyper. TETyper takes raw sequencing reads as input, and identifies: 1. Structural variation within a specific TE of interest, 2. Single-nucleotide variants (SNVs) within the TE, and 3. Flanking genetic context(s) of the TE. Variation within the TE captures signatures of micro-evolution, while the flanking sequences capture signatures of transposition, as every transposition event introduces a new genetic sequence context for the TE. This information can then be utilised for investigating transposition pathways, as well as gaining epidemiological insight in the context of resistance gene dissemination, both within outbreaks and globally. We demonstrate the utility of TETyper by applying it to a collection of over 3000 publicly available Illumina datasets containing *bla*<sub>KPC</sub>, and to IS26 within a clonal *K. pneumoniae* outbreak.

## THEORY AND IMPLEMENTATION

### Description of the TETyper pipeline

An overview of processing steps is shown in Fig. 1. Firstly, reads are mapped against a reference representing the TE of interest using bwa mem with default parameters [24]. For the remaining steps, only mapped reads are retained.



**Fig. 1.** Overview of processing steps in the TETyper pipeline.

To classify structural variation within the TE, we focus on deletions relative to the reference, since insertions and other rearrangements are difficult to classify reliably using short-read data. This is achieved by assembling the reads that map to the TE (along with any unmapped pairs) using spades [25], followed by BLASTn [26] to identify missing regions.

To identify SNVs, variant calling is performed under a diploid model using samtools mpileup [27], with variants excluded if they fall within a deleted region as identified above, or if they are not supported by at least one read in each direction. Heterozygous variants are assumed to represent within-sample mixtures (i.e. multiple, slightly different, copies of the TE), and are reported along with the number of reads supporting each nucleotide.

To identify flanking genetic context(s) of the TE, the user specifies the desired length of flanking sequence to classify,

which should be short enough that sequencing errors are rare. Reads mapping to the start/end of the TE sequence are examined, and the sequence of each read immediately adjacent to the start/end of the TE reference, of the length specified above, is extracted. To remove low-quality sequences, reads are discarded if the mapped sequence is too short (default threshold 30 bp), or if the minimum base quality within the flanking sequence is too low (default threshold 10). The resulting flanking sequences are then output if they pass additional thresholds for the number of supporting reads (defaults: 10 total, 1 each strand). In this way, if there is a single copy of the TE present in the sample, there should be a single flanking sequence identified at each end of the TE. On the other hand, every time a TE undergoes transposition, it inserts into a new genetic context. Therefore, if there are X copies of a TE in a sample, then there will be up to X distinct flanking sequences at each end of the TE. For

the applications below, we examine flanking sequences equal to TSD length (5 bp for Tn4401 and 8 bp for IS26), thus providing a simple signature for identifying shared insertion sites. For other applications, if the wider genetic context is of interest, then longer flanks can be used to obtain unique sequences for comparison with a reference genome.

Specific parameters used for running TETyper are described in Supplementary Methods.

## Validation

The TETyper output was validated using a subset of isolates from the datasets described below for which complete, closed references were previously generated using long-read sequence data [11, 28–30].

For the *bla<sub>KPC</sub>* dataset, we used 24 isolates representing a range of species, with different Tn4401 variants, and 1–2 Tn4401 copies each [11, 28, 29]. In all cases, the TETyper output was consistent with the Tn4401 elements present in the long-read assemblies (Table S2, available in the online version of this article).

Interestingly, for three of the six isolates with two Tn4401 copies (CAV1217, CAV1321 and CAV1741), the immediate Tn4401 flanking sequence was identical between the two copies, presumably because duplication of Tn4401 occurred via a mechanism other than Tn4401 transposition that also involved duplication of the surrounding sequence (e.g. via an additional flanking TE, as previously described for CAV1217 [28]). In these cases, TETyper only identified a single left/right flanking sequence and if this output were taken at face value, then the number of Tn4401 copies would be underestimated. This highlights the way in which TETyper's characterisation of the immediate flanking sequences is specifically designed to capture transposition of the reference TE, while other types of genetic mobility may be missed. In order to capture mobilisation events involving an additional flanking TE, TETyper could be used with the entire nested TE structure as a reference; however, this would require knowledge of the relevant TE structures involved.

For the other three isolates with two Tn4401 copies (CAV1042, CAV1392 and CAV1596), the immediate flanking sequences were different, and TETyper correctly identified two flanking sequences in each case. For one of these isolates (CAV1392), the coverage of the flanking sequences was sufficiently different that this information could be used to phase the left and right flanking sequences in the absence of long-read data.

For the IS26 analysis, there was one isolate that had previously been long-read sequenced (PMK1; accession CP008929.1–CP008933.1). There were six full-length and three truncated copies of IS26 in the long-read assembly, resulting in seven and eight left and right flanking genetic contexts respectively (Table S3). The flanking sequences identified by TETyper (Fig. 4) were a perfect match,

demonstrating the ability of TETyper to accurately identify flanking genetic contexts even in the presence of multiple TE copies.

It is worth noting that the TETyper output did not identify any deletions, even though truncated IS26 copies were present. This is expected, as when the multiple copies of IS26 are combined, there are no missing regions. Nevertheless, it is important to be aware that the method may be unable to capture structural variation when multiple copies are present.

On the other hand, one of the IS26 copies in the long-read assembly had a single SNV relative to the IS26 reference at position 107, resulting in a missense mutation in the IS26 transposase. However, TETyper did not identify any SNVs. This is a genuine discrepancy, as the heterozygous SNV output is designed to capture situations where multiple copies of the TE have SNV-level differences. However, in this case only one out of eight IS26 copies that cover the variant position has the mutation, and consistent with this the read pileup generated from mapping Illumina reads to the IS26 reference had only 12 % of reads at this position with the variant base. This highlights a limitation in the variant-calling algorithm used, as alternative alleles present at low frequency may not be detectable. Interestingly, all 34 samples showed similar frequencies of the variant base at this position, consistent with the variant allele being present across all samples, as indicated by the presence of the corresponding flanking sequences (ATTGTTTT/GGTCTTAA) in all samples (Fig. 4). For two samples (PMK21b and PMK25), the site was in fact called as heterozygous, indicating that the composition of IS26 elements within this sample collection sits close to the limit of detection for identifying heterozygous SNVs (as defined by the samtools variant-calling algorithm).

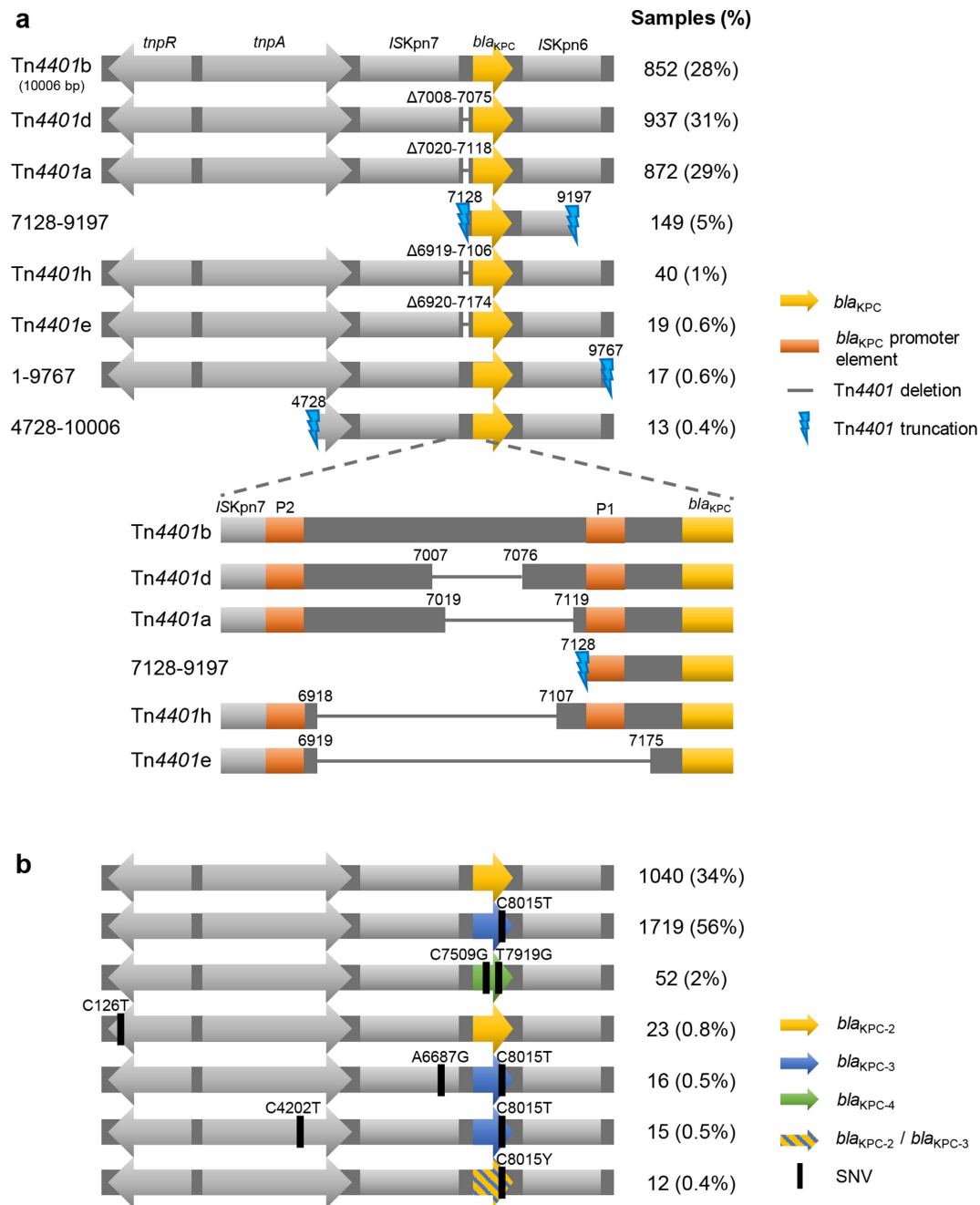
## Application to Tn4401

To demonstrate the utility of TETyper for epidemiological investigations, we applied it to 3054 *bla<sub>KPC</sub>*-positive Illumina datasets retrieved from a December 2016 snapshot of the European Nucleotide Archive [31] (Supplementary Methods, Table S1). The Tn4401b sequence from pKPC\_UVA01 (CP017937) was used as a reference.

### Structural variation in Tn4401

There were eight 'common' (found in  $\geq 10$  samples) structural variants of Tn4401 (Fig. 2a). The ancestral Tn4401b structure [32] was found in 852 out of 3054 (28 %) samples. Four variants represented different deletions immediately upstream of *bla<sub>KPC</sub>*, namely Tn4401d [33, 34], Tn4401a [32], Tn4401h [35] and Tn4401e [33, 34] in 937 (31 %), 872 (28 %), 40 (1.3 %) and 19 (0.6 %) samples respectively. The other three variants all represented truncations of Tn4401.

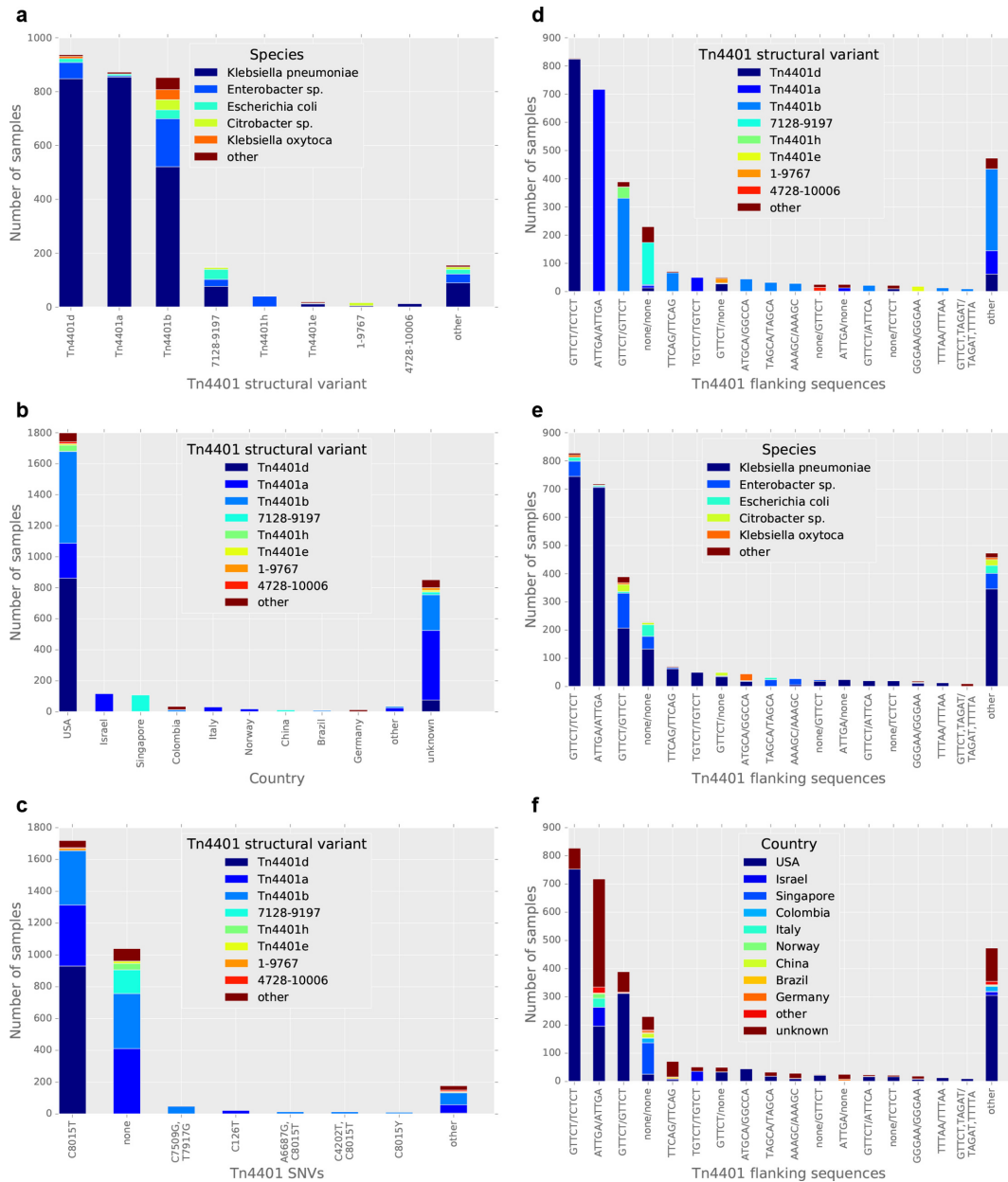
Deletions affecting the promoter region upstream of *bla<sub>KPC</sub>* have been shown to result in increased expression [34, 35], which is expected to be advantageous under antibiotic selection pressure. As several of these were observed, with no other common internal deletions across the 10 kb Tn4401



**Fig. 2.** Structure of Tn4401 showing common structural variants (a) and SNVs (b). Only variants found in at least ten samples are shown. For simplicity, SNV variants are all illustrated within a Tn4401b structural background.

sequence, this indicates that much of the structural variation observed may be due to selection rather than random genetic drift. Truncation of Tn4401 presumably prevents further transposition; one possible reason for the abundance of truncation variants is that they bring other TEs into the vicinity of *bla*<sub>KPC</sub> [36], thus providing alternative routes for gene mobilisation.

Specific structural variants were generally found in multiple host species, indicating wide horizontal dissemination via inter-species transfer (Figs 3a and S1a). Tn4401b was the most widely disseminated, being found in ten different genera, while Tn4401a was relatively restricted to *K. pneumoniae* (98%). Several different structural variants were present in samples from the



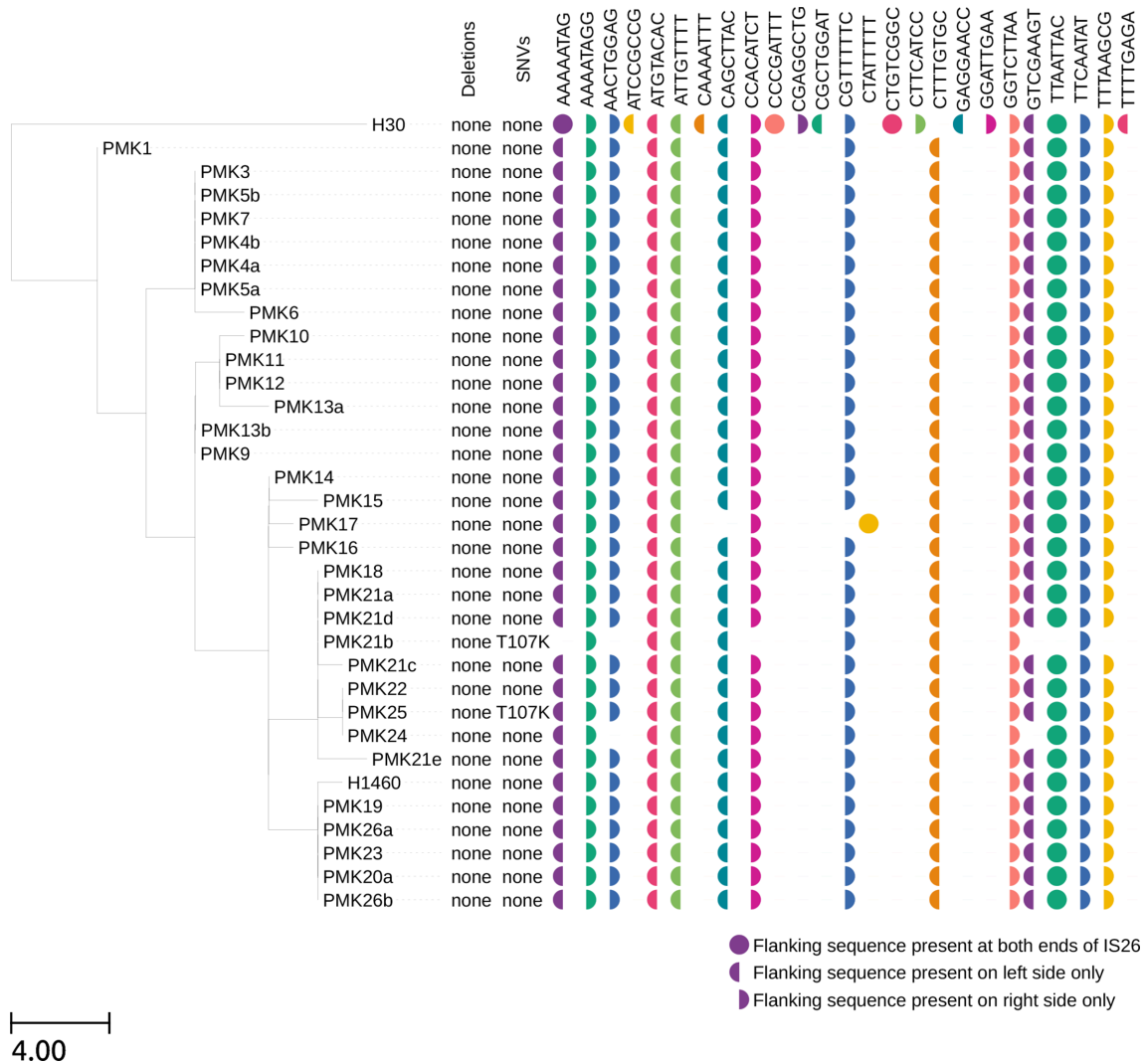
**Fig. 3.** Associations between Tn4401 structural variants, SNVs, flanking genetic contexts, host species, and countries of origin amongst a collection of 3054 *bla<sub>KPC</sub>* samples from the European Nucleotide Archive. (a) Distributions of host species for different structural variants of Tn4401. (b) Distributions of Tn4401 structural variants for different countries of origin. (c) Distributions of Tn4401 structural variants for different Tn4401 SNVs. (d–f) Distributions of Tn4401 structural variants (d), host species (e) and countries of origin (f) for different Tn4401 flanking genetic contexts.

USA, while samples from other countries generally had a single predominant variant (Figs 3b and S1b), supporting the origination and diversification of *bla<sub>KPC</sub>* and Tn4401 in Enterobacteriaceae in the USA. However, the dataset was heavily biased towards isolates from the USA, and for 851 out of 3054 (28 %) samples the country of origin was unknown, highlighting limitations in metadata availability.

### Single-nucleotide variation in Tn4401

Most SNV variation involved sites within the *bla<sub>KPC</sub>* gene (Fig. 2b), again implicating selection for KPC function in explaining observed variation. The two most common variants carried *bla<sub>KPC-2</sub>* and *bla<sub>KPC-3</sub>*, in 1040 out of 3054 (34 %) and 1719 out of 3054 (56 %) samples respectively, with each found in several different structural backgrounds (Figs 3c and S1c).





**Fig. 4.** Variation in IS26 amongst 34 ST15 *K. pneumoniae* isolates from an NDM-1 outbreak. TETyper output is annotated alongside a maximum likelihood phylogeny that was generated using IQ-TREE version 1.3.13 [38], after mapping to the MGH78578 reference as previously described [30]. Branch lengths are shown as SNVs per genome.

Interestingly, 12 samples showed polymorphism at the site that differentiates *bla*<sub>KPC-2</sub> and *bla*<sub>KPC-3</sub> (Fig. 2b), signifying a mixture of the two alleles and indicating that these samples most likely contain two copies of Tn4401, one with *bla*<sub>KPC-2</sub> and one with *bla*<sub>KPC-3</sub>. Minor allele percentages ranged from 10–49 %. These occurred in several Tn4401 structures, including Tn4401a, Tn4401b and Tn4401d, and several host species, including *Escherichia coli*, *K. pneumoniae*, *Klebsiella oxytoca* and *Enterobacter cloacae*, with varying flanking sequences. This indicates that the presence of multiple *bla*<sub>KPC</sub> variants may be a general phenomenon, indicating repeated multiple acquisition of *bla*<sub>KPC</sub> and/or repeated mutation converting *bla*<sub>KPC-2</sub> to *bla*<sub>KPC-3</sub> (or vice versa).

### Flanking genetic contexts of Tn4401

The most common 5 bp sequences flanking Tn4401 were GTTCT/TCTCT, ATTGA/ATTGA and GTTCT/GTTCT, present in 827 (27 %), 718 (24 %) and 389 (13 %) out of 3054 samples respectively (Figs 3d–f and S1d–f). ATTGA/ATTGA corresponds to the epidemic IncFII pKpQIL plasmid; these samples were almost exclusively Tn4401a-containing *K. pneumoniae*, but from a variety of geographic locations. GTTCT/GTTCT corresponds to Tn1/2/3-like elements (including Tn1331; see below), which have been described as containing Tn4401 in many different plasmid backbones [11]; these samples represented a wider variety of Tn4401 structures and host species. GTTCT/TCTCT is consistent with IncFIA pBK30661/pBK30683-like plasmids,

where Tn4401 is adjacent to a partial Tn1331 element on the left side only, presumably as a result of deletion on the right side following initial integration [37].

For 398 out of 3054 (13 %) samples, there was no flanking sequence identified on one or both sides of Tn4401, consistent with Tn4401 truncation. These truncated Tn4401 structures are presumably not transpositionally active, so TETyper would be unable to capture ongoing mobilisation events in these cases. For the vast majority, however, Tn4401 appeared to be intact; 2368 (78 %) samples had a single flanking sequence identified on each side of Tn4401, indicating a single intact copy, and 288 (9 %) had multiple flanking sequences on one or both sides, indicating multiple copies.

Of those with a single copy, 1417 out of 2368 (60 %) had the same 5 bp sequence at both ends of Tn4401, consistent with TSD following standard intermolecular transposition. Surprisingly, 951 out of 2368 (40 %) showed different 5 bp sequences, indicating disruption of TSD signatures and suggesting that intramolecular transposition of Tn4401 or other rearrangements disrupting TSDs may be relatively common.

Altogether, there were 193 and 213 distinct 5 bp sequences flanking the left and right sides of Tn4401 respectively, and a total of 272 distinct flanking sequence profiles, indicating relatively frequent transposition. This diversity indicates that the classification of flanking genetic contexts in this way may be useful for epidemiological tracking by providing higher genetic resolution than strain typing alone.

## Application to IS26

To demonstrate the utility of TETyper for analysing specific TE mobility events without relying on a reference genome, we applied it to IS26 for 34 closely related *K. pneumoniae* ST15 isolates from an NDM-1 outbreak in Nepal [30]. These isolates varied in the number and sequence of genetic contexts of IS26, with evidence for 4–14 copies per isolate (Fig. 4, Table S1). In some cases, the TETyper output provided higher genetic resolution than a standard phylogenetic approach, with IS26 flanking sequence profiles differing between pairs of isolates with 0 chromosomal SNVs (Fig. 4; PMK21b vs PMK18/21a/21d and PMK24 vs PMK22/25).

## Conclusion

We have developed a novel bioinformatic pipeline, TETyper, for classifying sequence variation and flanking genetic contexts of TEs from short-read WGS data, without requiring a genome reference. We have demonstrated the utility of TETyper by applying it to Tn4401 for a large, global *bla*<sub>KPC</sub> collection, as well as IS26 within a small, defined outbreak. This revealed surprising diversity in both cases, and provided insights into patterns of transposition and mutational change within Tn4401. In an epidemiological context, the within-TE variation and transposition signatures identified by TETyper could be used to facilitate

higher-resolution resistance gene tracking related to gene mobility than is currently possible using other WGS-based methods.

## Funding information

The research was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England (PHE), in collaboration with the University of Virginia, and by a contract from the US Centers for Disease Control and Prevention (CDC) Broad Agency Announcement (BAA 2016-N-17812). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health, Public Health England or the CDC. N. S. is supported by a Public Health England/University of Oxford Clinical Lectureship.

## Acknowledgements

We thank Zamin Iqbal and Phelim Bradley for assistance with BIGSI. We also thank the HPRU Steering Group.

## Conflicts of interest

The authors declare that there are no conflicts of interest.

## Ethical statement

No experiments involving humans or animals were performed for this study.

## Data bibliography

1. TETyper source code, Sheppard AE, GitHub, (2018).

## References

1. Carattoli A. Plasmids and the spread of resistance. *Int J Med Microbiol* 2013;303:298–304.
2. Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. Everyman's guide to bacterial insertion sequences. *Microbiol Spectr* 2015;3:MDNA3-0030-2014.
3. Partridge SR. Resistance mechanisms in *Enterobacteriaceae*. *Pathology* 2015;47:276–284.
4. Harmer CJ, Moran RA, Hall RM. Movement of IS26-associated antibiotic resistance genes occurs via a translocatable unit that includes a single IS26 and preferentially inserts adjacent to another IS26. *MBio* 2014;5:e01801-14.
5. Poirel L, Lartigue MF, Decusser JW, Nordmann P. ISEcp1B-mediated transposition of *bla*<sub>CTX-M</sub> in *Escherichia coli*. *Antimicrob Agents Chemother* 2005;49:447–450.
6. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med* 2014;2:285–292.
7. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A et al. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *N Engl J Med* 2013;369:1195–1205.
8. Senn L, Clerc O, Zanetti G, Basset P, Prod'homme G et al. The stealthy superbug: the role of asymptomatic enteric carriage in maintaining a long-term hospital outbreak of ST228 methicillin-resistant *Staphylococcus aureus*. *MBio* 2016;7:e02039.
9. Kwong JC, Lane CR, Romanes F, Gonçalves da Silva A, Easton M et al. Translating genomics into practice for real-time surveillance and response to carbapenemase-producing *Enterobacteriaceae*: evidence from a complex multi-institutional KPC outbreak. *PeerJ* 2018;6:e4210.
10. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* 2017;3:e000128.
11. Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A et al. Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene *bla*<sub>KPC</sub>. *Antimicrob Agents Chemother* 2016;60:3767–3778.



12. 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.
13. Munoz-Price LS, Poirel L, Bonomo RA, Schwaber MJ, Daikos GL *et al.* Clinical epidemiology of the global expansion of *Klebsiella pneumoniae* carbapenemases. *Lancet Infect Dis* 2013;13:785–796.
14. Lee CR, Lee JH, Park KS, Kim YB, Jeong BC *et al.* Global dissemination of carbapenemase-producing *Klebsiella pneumoniae*: epidemiology, genetic context, treatment options, and detection methods. *Front Microbiol* 2016;7:895.
15. Bonomo RA, Burd EM, Conly J, Limbago BM, Poirel L *et al.* Carbapenemase-producing organisms: a global scourge. *Clin Infect Dis* 2018;66:1290–1297.
16. Mathers AJ, Peirano G, Pitout JD. The role of epidemic resistance plasmids and international high-risk clones in the spread of multi-drug-resistant *Enterobacteriaceae*. *Clin Microbiol Rev* 2015;28:565–591.
17. Zhang Y, Wang Q, Yin Y, Chen H, Jin L *et al.* Epidemiology of carbapenem-resistant *Enterobacteriaceae* infections: report from the China CRE network. *Antimicrob Agents Chemother* 2018;62.
18. Correa A, del Campo R, Perenguez M, Blanco VM, Rodríguez-Baños M *et al.* Dissemination of high-risk clones of extensively drug-resistant *Pseudomonas aeruginosa* in Colombia. *Antimicrob Agents Chemother* 2015;59:2421–2425.
19. He S, Hickman AB, Varani AM, Siguier P, Chandler M *et al.* Insertion sequence IS26 reorganizes plasmids in clinically isolated multidrug-resistant bacteria by replicative transposition. *MBio* 2015;6:e00762.
20. Hawkey J, Hamidian M, Wick RR, Edwards DJ, Billman-Jacobe H *et al.* ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics* 2015;16:667.
21. Nakagome M, Solovieva E, Takahashi A, Yasue H, Hirochika H *et al.* Transposon Insertion Finder (TIF): a novel program for detection of *de novo* transpositions of transposable elements. *BMC Bioinformatics* 2014;15:71.
22. Gilly A, Etcheverry M, Madoui MA, Guy J, Quadrana L *et al.* TE-Tracker: systematic identification of transposition events through whole-genome resequencing. *BMC Bioinformatics* 2014;15:377.
23. Nelson MG, Linheiro RS, Bergman CM. McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3* 2017;7:2763–2778.
24. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.
25. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.
27. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27:2987–2993.
28. Mathers AJ, Stoesser N, Chai W, Carroll J, Barry K *et al.* Chromosomal integration of the *Klebsiella pneumoniae* carbapenemase gene, *bla<sub>KPC</sub>*, in *Klebsiella* species is elusive but not rare. *Antimicrob Agents Chemother* 2017;61:e01823–16.
29. Sheppard AE, Stoesser N, Sebra R, Kasarskis A, Deikus G *et al.* Complete genome sequence of KPC-producing *Klebsiella pneumoniae* strain CAV1193. *Genome Announc* 2016;4:e01649–15.
30. Stoesser N, Giess A, Batty EM, Sheppard AE, Walker AS *et al.* Genome sequencing of an extended series of NDM-producing *Klebsiella pneumoniae* isolates from neonatal infections in a Nepali hospital characterizes the extent of community- versus hospital-associated transmission in an endemic setting. *Antimicrob Agents Chemother* 2014;58:7347–7357.
31. Bradley P, den Bakker H, Rocha E, McVean G, Iqbal Z. Real-time search of all bacterial and viral genomic data. *bioRxiv* 2017.
32. Naas T, Cuzon G, Villegas MV, Lartigue MF, Quinn JP *et al.* Genetic structures at the origin of acquisition of the  $\beta$ -lactamase *bla<sub>KPC</sub>* gene. *Antimicrob Agents Chemother* 2008;52:1257–1263.
33. Kitchel B, Rasheed JK, Endimiani A, Hujer AM, Anderson KF *et al.* Genetic factors associated with elevated carbapenem resistance in KPC-producing *Klebsiella pneumoniae*. *Antimicrob Agents Chemother* 2010;54:4201–4207.
34. Naas T, Cuzon G, Truong HV, Nordmann P. Role of *ISKpn7* and deletions in *bla<sub>KPC</sub>* gene expression. *Antimicrob Agents Chemother* 2012;56:4753–4759.
35. Cheruvanky A, Stoesser N, Sheppard AE, Crook DW, Hoffman PS *et al.* Enhanced *Klebsiella pneumoniae* carbapenemase expression from a novel Tn4401 deletion. *Antimicrob Agents Chemother* 2017;61:e00025–17.
36. Stoesser N, Sheppard AE, Peirano G, Anson LW, Pankhurst L *et al.* Genomic epidemiology of global *Klebsiella pneumoniae* carbapenemase (KPC)-producing *Escherichia coli*. *Sci Rep* 2017;7:5917.
37. Chen L, Chavda KD, Melano RG, Hong T, Rojzman AD *et al.* Molecular survey of the dissemination of two *bla<sub>KPC</sub>*-harboring IncFIA plasmids in New Jersey and New York hospitals. *Antimicrob Agents Chemother* 2014;58:2289–2294.
38. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.

### Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at [microbiologyresearch.org](http://microbiologyresearch.org).